Volume 6, Number 3, 2019

REGION

The Journal of ERSA
Powered by WU

Table of Contents

Funded by

REGION, The journal of ERSA / Powered by WU
ISSN: 2409-5370

ERSA: http://www.ersa.org
WU: http://www.wu.ac.at

Editorials

REGION
The Journal of ERSA
Powered by WU

Volume 6, Number 4, 2019, E1–E2
DOI: 10.18335/region.v6i4.309

journal homepage: region.ersa.org

Fueling Research Transparency: Computational Notebooks and the Discussion Section

Sierdjan Koster[1], Francisco Rowe[2]

[1] University of Groningen, Groningen, The Netherlands
[2] University of Liverpool, Liverpool, United Kingdom

Received: February 18 2020/Accepted: March 4 2020

The results of academic research, often publicly funded, should be easily available for a wide audience that includes fellow researchers, policy makers, journalists and anyone who takes an interest. On top of this, the research itself should be done in a transparent way so that results can be reproduced or perhaps falsified, as transparent research is good practice across all academic disciplines. Arguably, however, it is even more central in fields such as Regional Science with salient and tangible implications for local, regional, national and international socio-economic policies. The advent of internet and the growing capacity of computers has increased both the possibilities and the societal demand for transparent science.

Accommodating transparent research in Regional Science is a cornerstone of REGION. By its very nature as an online and open access journal, the contributions are made available to everyone. As it does not charge any submission fees, the journal is inclusive and invites contributions from across the globe. Also, in the longstanding and successful Resources section, REGION aims to unlock available datasets, visualization techniques and empirical approaches for a large audience. We are now proud to present two new types of publication that further accommodate transparent research in Regional Science: The Discussion section, and Computational Notebooks (see Figure 1).

The Discussion section aims to accommodate contributions that reflect on Regional Science as a field of science. Such reflections can pertain to fields of study within regional science, research agendas, but also to the organization of the field including the position of young scientists, gender representation and geographical inclusiveness of Regional Science. We invite contributions that put forward ideas for the further development of the field, but also retrospective accounts are very welcome. This issue has the first item in the Discussion section, which puts forward a research agenda on Transport research that emanates from the NECTAR network (Geurs, Macharis 2019).

Computational notebooks aim to augment the replicability, reproducibility and reach of research in Regional Science. Computational notebooks allow one to present analyses and integrate code, specialised software, dependencies, results and descriptive text into a single 'computational narrative' to be shared, read and executed by others. The interactive and narrative nature of computational notebooks provides unique opportunities for sharing computational research, enabling reproducibility in published scientific research. This is particularly important at a time of increased complexity of scientific studies. Computational notebooks are also valuable vehicles for teaching and demonstration of analytical tools. They can augment the impact of research beyond its primary academic objectives by extending original analysis and by reaching non-academic communities interested in Regional Science. The interactivity of notebooks can engage policy makers and the general public in ways that standard academic journal publications cannot.

(a) Discussion Section (b) Computable Notebooks

Figure 1: New publication types in REGION

REGION has recognised the potential of computational notebooks and the changing publication landscape, and we now facilitate the publication of computational notebooks, both as part of regular article submissions and as dedicated publications for the Resources section. We are proud to present the first contributions with computational notebooks in this issue (Comber 2020, Boeing 2020, Patias 2020) and we invite our contributors to incorporate computational notebooks in their submissions. In the near future, we will publish a dedicated special issue on the use of notebooks in the Regional Science.

We are confident that with these innovations, REGION can continue to play an active role in making Regional Science into a transparent and reproducible science.

References

Boeing G (2020) Urban street network analysis in a computational notebook. *REGION* 6[3]: 39–51. CrossRef.

Comber S (2020) Demonstrating the utility of machine learning innovations in address matching to spatial socio-economic applications. *REGION* 6[3]: 17–37. CrossRef.

Geurs K, Macharis C (2019) The future of european communication and transportation research: a research agenda. *REGION* 6[3]: D1–D21. CrossRef.

Patias N (2020) Exploring long-term youth unemployment in europe using sequence analysis: a reproducible notebook approach. *REGION* 6[3]: 53–69. CrossRef.

Articles

REGION
The Journal of ERSA
Powered by WU

Volume 6, Number 3, 2019, 1–16
DOI: 10.18335/region.v6i3.249

journal homepage: region.ersa.org

Business travel decisions and high-speed trains: An ordered logit approach

Federica Rossi[1], **Rico Maggi**[2]

[1] Politecnico di Milano, Milan, Italy
[2] Università della Svizzera italiana, Lugano, Switzerland

Received: 25 July 2018/Accepted: 15 November 2019

Abstract. The paper studies the potential impact on business travel of the new high-speed railway line project, called AlpTransit, which will link Lugano, the small economic hub of the southern part of Switzerland, with Zurich, one of the major Swiss economic centres, situated north of the Alps. This infrastructure has enabled travel time between the two cities to decrease considerably from about three hours to less than two hours by the end of 2020.

The question that we pose in this paper is what impact high-speed trains could have, in the short to medium term, on business travel between the two hubs (ex-ante evaluation). Indeed, given the travel time, firms could increase their business-to-business one-day trips, boosting face-to-face interactions within and among enterprises. Our curiosity more specifically regards the potential impact of the change in travel time on the propensity to travel of employees with different functions in various types of firms.

An online survey was conducted among firms located in Ticino, the Swiss Canton that includes Lugano. The data is analysed using four ordered logit models, one for each employee category (CEO, administrative staff, sales personnel, specialists), since hierarchical position and professional status influence business travel characteristics. Results show that internal firm characteristics, such as sector, frequency and destination of current business travels significantly influence the propensity to travel to Zurich more often due to AlpTransit.

JEL classification: D22, R40

Key words: High-speed train, business travel, ordered logit, firm behaviour

1 Introduction

One of the persistent analytical issues in the economic geography of transport relates to the assessment of the contribution that transport infrastructure makes to the economy.

The majority of the literature analyses transport impact on regional development (Gutiérrez 2001, Marti et al. 2007, Carbo et al. 2019) and companies' relocation strategies (Leitham et al. 2000, Kawamura 2004, De Bok, Sanders 2005), and therefore focuses on long term effects. However, by reducing travel time, a new infrastructure also has various impacts in the short and medium term, in particular on daily choices of firms.

The paper presents an empirical analysis on firms located in Ticino, the southern Swiss Canton, by studying the influence on one-day business-to-business trips of the new high-speed railway line project, called AlpTransit, which will link Lugano, a small Swiss city in the south, with Zurich, the Swiss economic capital situated north of the Alps. This infrastructure is an ambitious railway project: its first and main segment (Gotthard base tunnel) has been opened in late 2016[1], and the Monte Ceneri base tunnel will complete the North-South link by 2020. Due to the new high-speed railway line, the reduction in travel time will be very significant: from three hours to less than two between the cities of Zurich and Lugano. The Swiss population, and in particular firms' managers, are aware of this project, because of the large advertising campaign that has been conducted.

More precisely, the goals of this work are to investigate if and how high-speed trains could change firms' business relations between a small city (Lugano) and a dominant one (Zurich), which have similar economic structures (they are both specialised in business services, in particular financial services). Second, this work aims to understand if a significant reduction in travel time will have a differentiated influence on some specific firms. Finally, this work will identify the link between firms' characteristics and future business trips within various employee categories.

In order to achieve these goals, the probability that face-to-face contacts will increase after AlpTransit is analysed by distinguishing employee categories: CEO and upper management, administrative staff, sales personnel and specialists. This specification helps identify the probable meeting purpose, for example visiting clients, branches, government departments or attending courses, fairs, conferences and conventions, opening or closing new units, projects, R&D, etc. (Swarbrooke, Horner 2001, Welch, Worm 2005, Beaverstock et al. 2009). Indeed, strategic decisions "travel" with the CEO and upper management (Jones 2007) and the areas where face-to-face contacts are fundamental for the success of companies are sales and business development (HBR 2009). This specification is also motivated by the fact that not all workers travel: professional status and hierarchical position are significant factors, which influence the business travel characteristics (Aguilera 2008).

While most of the past studies are ex-post evaluations of the impact of an infrastructure, this paper offers an ex-ante evaluation of the AlpTransit project, using individual firm level data. In order to do this, we rely on the stated preference methods, which allow taking into account hypothetical behaviours in the future.

The paper is organised as follows: the next section is dedicated to an overview of the literature, while in the third section the theoretical framework and hypothesis are presented. The following two parts concern the survey and the empirical model. The sixth section highlights the main results and a brief discussion of them. Finally, some conclusions and future research ideas are presented.

2 Literature review

Nowadays, doing business means facing the increasing interconnection and globalisation of our world. In this context, business travel (BT) has become a quite common and diffuse practice (Swarbrooke, Horner 2001, Aguilera 2008, UNWTO 2012), and an essential feature of globalised trends like outsourcing, spatial specialisation and multi-plant companies.

Business visits are defined as work-related trips to an irregular place of work, lasting less than 12 months (Aguilera 2008). These kinds of trips have continuously grown in number, despite the increasing availability of other forms of distance communication, such as video conferencing (Choo et al. 2007). Moreover, BT expands the market potential of firms, creating new opportunities for acquiring contracts (Blum et al. 1997, Jones 2007), improves global corporate productivity, facilitates the creation of new jobs, and attracts new clients. Further, BT increases profits, sales, partnerships and innovation (Beaverstock et al. 2009, Machikita, Ueki 2010, WTTC 2011, Gustafson 2012). Two recent econometric studies provide evidence on the impacts of business travel. Poole (2010) demonstrates that business travel has a positive impact on the extensive export margin, thereby helping to overcome informational asymmetries in international trade. Concerning innovation,

[1] Note that the survey described in the fourth paragraph was implemented in 2014.

Hovhannisyan, Keller (2015) show that BT has a significant effect beyond technology transfer, and in particular, leads to an increase in patenting. The literature thus provides ample evidence of business travel as an effective way of transferring knowledge, conserving long-term relationships, and coordinating and monitoring, all of which have a direct impact on trade and offshoring activities (HBR 2009, Cristea 2011).

Another strand of literature relates more closely to our topic, it concerns accessibility improvements caused by the introduction of high-speed train connections. In the economic literature, a lot has been written about the impact of high-speed trains (HST). For example, due to HST, isolated markets have been better integrated, monopolistic positions have been reduced and competition as well as productivity have increased (Blum et al. 1997). Moreover, there is an open debate on the impact of HST on GDP (Banister, Thurstain-Goodwin 2011) and on regional development (Vickerman et al. 1999, Gutiérrez 2001). The literature has also highlighted that this type of infrastructure will create new opportunities by offering additional locational advantages for economic activities and commuting/business trips, thus changing the traditional city roles (Ureña et al. 2009).

Looking at some European experiences, we can generally find an ex-post positive impact of high-speed trains on BT. In the Lille-Paris line, provided by the new TGV, business travel has increased by 1/3 in both directions (Vickerman, Ulied 2009). This is also true for the Lyon-Paris line, where there is a growth in face-to-face contact in both directions, principally due to the activities between the subsidiary offices and the headquarters (Harman 2006). In general, many intermediate cities that are between two or more metropolitan areas -for example Lille, Cordoba, Zaragoza and Lyon- have numerous advantages from the building of new high-speed infrastructures. Thus, these intermediate cities experience a flourishing of their service sector (Ureña et al. 2009). Moreover, HST positively influences the city's position in the European urban hierarchy, empowering both dominant cities, i.e. those with the highest rank-order, and intermediate cities (McCann 2001, Mazzeo 2012). This urban structure replicates the Swiss situation well: due to high-speed trains, Lugano, a small intermediate city, will be closer to Zurich, a dominant economic pole.

The impacts of transport improvements on business travel on a micro level, which are the focus of this paper, have found less attention in research. According to Gutiérrez (2001), HST improves accessibility to cities by shortening travel time, and in particular, according to the definition of inbound/outbound accessibility (Törnqvist 1984), increases the face-to-face contact opportunities during a one-day trip, especially business-to-business trips (Blum et al. 1997, Willingers et al. 2007). The relevant literature underlines that accessibility is fundamental both for the decisions that a company takes and for the company's performance. In this sense, high-speed trains not only create possibilities for more face-to-face contacts with other enterprises, but also with customers, suppliers, partners and their workforce (Blum et al. 1997, Bruinsma, Rietveld 1998). The majority of such studies so far have been qualitative analyses, with no econometric measurements of the relationship between the increase in business travel and HST and have ambiguous results (Blum et al. 1997, Kobayashi, Okumura 1997, Harman 2006, Willingers et al. 2007). This is likely due to the scarce availability of micro data directly collected from firms.

Literature on the evaluation of travel time savings for business travellers can contribute to the arguments in this paper. Wardman et al. (2013), in a report commissioned by the UK Department for Transport, provide a comprehensive overview on methods and empirical evidence for the evaluation of travel time savings of business travellers. For Switzerland, Axhausen et al. (2006) estimate distance and income dependent values of travel time savings for various trip purposes and find a value of 30 CHF/hour for business travel. We will draw partly on this literature in our theoretical considerations below.

Concerning the rationale for increasing business travel as a reaction to a reduced travel time, we will draw on the following works: Maggi 1989; Button et al. 1993; Button and Maggi 1995. These papers treat the issue from the perspective of choice between traveling for a face-to-face meeting and telecommunication contacts. The basic rationale is on the one hand, the superiority of face-to-face communication in terms of the quality of interaction, and on the other hand, the added transport cost.

Hugoson (2001) also argued that the choice of face-to-face business contacts is associated with non-standardised information exchange and high transaction costs. These costs affect a firm's profit and differ between forms of interaction (i.e. telecommunication vs face-to-face contacts). The face-to-face meeting is chosen if the expected added profit from meeting is greater than the expected profit of a mediated contact and greater than the expected profit of no contact at all.

Apart from travel time, our theoretical framework (discussed in the next paragraph) does not provide indications for the specification of our empirical model. We therefore rely on empirical literature, which will now be reviewed.

As previous studies suggest, the variables influencing business travel behaviour can be grouped into characteristics of firms and current business relations.

Above all, the firm's sector influences the perception of the importance of accessibility. In particular, the empirical evidence shows that in the tertiary sector, e.g. business services and R&D, face-to-face contacts are essential for the success of the activities (Blum et al. 1997, Jones 2007, Aguilera 2008). For this reason, the business services sector is the one that could benefit more from HST (Harman 2006). While a firm's size is a determinant for the perception of accessibility, we have not found any evidence supporting the hypothesis that the age of a firm significantly influences travel behaviour.

Another important determinant is the firm's spatial organisational structure, in terms of subsidiaries, branches and headquarters. In this context, business travel is essential to tying together spatially distributed subsidiaries (Aguilera 2008, Beaverstock et al. 2009).

Concerning current business relations, interactions with clients and suppliers are one of the main reasons for face-to-face contacts (Aguilera 2008, Beaverstock et al. 2009, Cristea 2011). According to a survey by the Harvard Business Review, more than 89% of review subscribers agree that face-to-face contacts are essential to "sealing a deal".

The destination of current business trips and the means of transport used can also influence the future behaviour (indeed, travel behaviour is based on habits; see Aarts et al. 1998, Aarts, Dijksterhuis 2000). We can imagine that firms, which nowadays have partners in a specific destination and visit them, will be the first to take advantage of new faster infrastructure.

3 Theoretical framework and hypothesis

As highlighted above, business travel is an essential ingredient of a company's activity in the context of globalisation, specialisation and spatial distribution of activities. Every firm interacts with its internal and external network of clients, suppliers etc. through a certain number of contacts. These contacts can take either the form of face-to-face meetings or any kind of telecommunication interaction. For simplicity, we assume the total number of business contacts to be fixed for a firm, and equal to $\bar{M}$. This is theoretically equivalent to a conditional demand for face-to-face meetings and telecommunication contacts, respectively. Therefore, following Maggi (1989), we model this decision as a cost-minimisation strategy regarding the number of direct vs telecommunication contacts. For the modelling of an unconditional demand for business meetings see Hugoson (2001). In our survey, it was not feasible to ask for the number of telecommunication contacts, and hence we observe only the current and future direct contacts. The future direct contacts might also include face-to-face contacts, which are stimulated by the new connection and thus increase the overall number of contacts. We will use the term "meetings" for contacts, where M_{ff} are face-to-face meetings and M_{tc} are telecommunication meetings. These meetings will in reality take many forms, concern various contents, and convey messages with different degrees of complexity, as their content can range from negotiations to simple internal briefings among departments.

We develop our argument defining the cost curves for face-to-face and telecommunication meetings, respectively. The differences in cost between the two modes regard on the one hand the presence of travel time (T_{travel}) which is added to meeting time, and of monetary travel cost (C_{travel}) only for face-to-face contacts. On the other hand is the cost dependence on complexity of content, only for telecommunication.

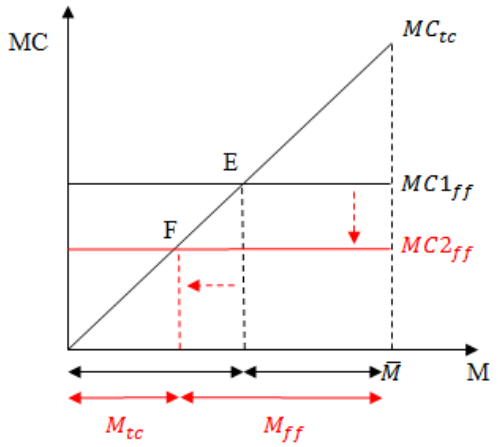


Figure 1: Marginal costs of face-to-face and telecommunication meetings

The two cost functions are:

$$\begin{aligned} C_{ff} &= VoT * (T_{travel} + T_{meeting}) * M + C_{travel} * M \\ C_{tc} &= VoT * T_{meeting} * f(M) \quad \text{with} \quad f'(M) > 0 \quad \text{and} \quad f''(M) > 0 \end{aligned}$$

Where:

C_{ff} ... Cost for face-to-face meetings,
C_{tc} ... Cost for telecommunication meetings,
VoT ... Value of time,
T_{travel} ... Travel time,
$T_{meeting}$ ... Meeting time,
C_{travel} ... Monetary travel cost,
M ... Number of meetings,
$f(M)$ ... function, which reflects the complexity of content in the meeting.

The intuition behind the formulation of the cost function for telecommunication is simply that the marginal cost of transmitting increasingly complex contents over a telecommunication mode is increasing. Standardising for simplicity meeting time to one ($T_{\text{meeting}} = 1$) and excluding corner solutions, cost-minimisation implies that the marginal costs of the two types of meetings are equal $MC_{ff} = MC_{tc}$ where:

$$\begin{aligned} MC_{ff} &= VoT * (T_{travel} + 1) + C_{travel} \\ MC_{tc} &= VoT * f'(M) \end{aligned}$$

Just as an example, if $f(M) = M^2$, the marginal cost for telecommunication meetings is linear, as illustrated in Figure 1. In this figure, the meetings are ordered by increasing complexity from left to right on the horizontal axis, and marginal costs on the vertical axis. The total number of meetings $\bar{M}$ is distributed between a certain number of face-to-face meetings and telecommunication interactions. When a decrease in travel time occurs, the line of the marginal cost for face-to-face contacts moves down (Figure 1 in red) and provokes a substitution of some of the telecommunication meetings of a certain complexity with face-to-face meetings. This substitution effect is realistic, given the experience that face-to-face contacts are the norm for very small local distances and in-house meetings.

Note, we do not consider a variation in the monetary travel cost. While in general it can be assumed that prices might increase if the stakeholders of an infrastructure project want to recover their investment, this is not the case of the entirely publicly funded AlpTransit project. In Switzerland, the ticket prices of public transports are only distance dependent and, in order to improve accessibility, the standard policy is to invest in a new infrastructure, without adjusting prices.

Overall, according to these theoretical considerations, we expect that AlpTransit will have a positive impact on business trips, since it will reduce travel times.

4 The survey

The data was collected with an online survey among a sample of firms located in Ticino in 2014. The questionnaire was explicitly addressed to the CEO/upper management of the firm.

In order to build the sample, we used the Bureau van Dijk business listing, which reported some characteristics of the firms, such as sector, size, financial data, etc. Among them, sector is the only characteristic reported for all firms, while the other characteristics often are absent or not updated. Therefore, we stratified our sample based on sector. In 2014, the overall number of firms located in Ticino was divided as follows: 8.4% in primary sector, 14.4% in secondary sector, and 77.3% in tertiary sector. If we exclude the primary sector[2], the percentages increase to 16% for the secondary sector, and to 84% for the tertiary sector. We take these values as references in the sampling strategy. Therefore, a sample of 5890 firms was randomly selected from the Bureau van Dijk business listing, reflecting the sectoral composition described above.

The overall response rate was 13%, in line with previous studies using questionnaires with firms: Kawamura (2004) obtained a response rate of 12.1% in a study on perceived accessibility and location choice of firms in the Chicago region (USA). Targa et al. (2006) obtained a response rate of only 6% in a survey on firms' relocation decisions in the state of Maryland (USA).

Obtaining high response rates in surveys with firms is complex, since the completion of the questionnaire requires that the survey is addressed to the right person within the firm, and that the CEO / upper management find time to complete it. Takin this into consideration, our response rate seems reasonable.

The response rate is slightly different between the two strata: 18% for the secondary sector and 12% for the tertiary sector. Looking at the differences between respondents and non-respondents, we found that 24.6% of non-respondents are in the secondary sector (versus 27.6% of respondents) and 75.4% of non-respondents belong to the tertiary sector (versus 72.4% of respondents).

From the 773 questionnaires filled in, we exclude incomplete questionnaires and all firms with zero employees (which are mainly foreign branches with only legal residence in Ticino), since they are not relevant for the purposes of our analysis. Therefore, 696 firms are considered in the model.

The survey was built following the most recent examples in the literature and can be divided into four main sections. The first section aims to collect general information about the enterprise: sector, age, spatial organisation (head office and branches), size, as well as information on relevant geographic markets (where suppliers and customers are located). In the second section, firms are asked to rate the importance and presence in Ticino of some location factors, such as accessibility, cost of labour, business taxes, etc.

The third part is devoted to business relations, in particular to understanding frequency, destination and means of transport used for business trips, with a focus on trips to Zurich. In the questionnaire, we specify that business visits are work-related travels to an irregular place of work. Finally, the last section contains some hypothetical questions on future actions of companies and in particular possible relocation outside Ticino and changes in BT due to AlpTransit.

Considering the possibility of easily reaching the train station, the geographical distribution of firms is important. The majority of respondents are concentrated in the

[2]Given the subject of our analysis (business travel), this omission is plausible.

Table 1: Descriptive statistics of respondents

Category		Number	%
Sector	Secondary	192	27.6
	Tertiary	504	72.4
Age	Less than 3 years old	180	25.9
	More than 4 years old	516	74.1
Organizational structure	Sole proprietorship	546	78.4
	Branch	80	11.5
	Headquarter	70	10.1
Size	Micro	495	71.1
	Small	161	23.1
	Medium-Large	40	5.8

city of Lugano, which is the main economic pole of Ticino and is the financial centre of the canton. The cities of Mendrisio and Chiasso (both in the south) follow in terms of highly populated areas. This concentration in the southern part of the canton, near Lugano where AlpTransit will stop, confirms that our respondents set is appropriate for the analysis.

Table 1 shows some descriptive statistics regarding the respondents to the questionnaire: 72.4% of firms operate in the tertiary sector; they are mainly firms, which are more than four years old (74.1%) and sole proprietorship (78.4%). Moreover, about 71% of respondents are micro firms (with less than 10 workers). The sample respondents' characteristics, which seem quite unbalanced, actually reflect the economic structure of Ticino, which is mainly composed of micro firms (86% of the businesses in Ticino in 2014) and firms operating in services (77.3%).

5 The empirical model

As stated above, the last part of the survey contains some hypothetical questions, formulated as Likert scales, on a firms' future behaviour and in particular on possible changes in business travel due to AlpTransit. At the time of the inquiry in 2014, the new connection had been under construction for 15 years and still had 2 more years until completion, i.e. start of train operations. All train schedules, in terms of travel times, frequencies and train stops had already been defined by the Swiss Federal Railways. Therefore, we have no concerns regarding possible strategical answering by respondents.

We use the question, "How likely is it that the following employee categories (CEO, sales personnel, administrative staff and specialists) will travel more to Zurich, due to AlpTransit? 1 (Not at all likely), 5 (Very likely or almost certain)" as an indicator for the dependent variable in the ordered logit models.

The distribution of the propensity to increase BT to Zurich across the employee categories is shown in Table 2. Almost 14% of the respondents claim that their CEOs will travel more to Zurich due to AlpTransit (including both "likely" and "almost certain" categories). This percentage decreases for the other employee categories: respectively to 4.7% for administrative staff, to 10.4% for sales personnel and to 7.1% for specialists. Using these descriptive statistics, we have preliminary evidence that the considered employee categories have a different propensity to increase their travel due to AlpTransit.

Ordered logit models are used when the dependent variable is ranked on a scale. In particular, they are applied in stated preference choice experiments with Likert scales (Greene, Hensher 2009, Hess, Daly 2014).

Ordered logit models allow an analysis of hypothetical situations. Using self-assessed evaluations, we could try to capture agent's behaviour in a future situation. Many applied works, on various topics, can be found in the literature, for example: education (Machin,

Table 2: Frequency of the propensity to increase BT to Zurich due to AlpTransit (%)

	1 – Not at all likely	2 – Unlikely	3 – Neither unlikely nor likely	4 – Likely	5 – Very likely or almost certain
Increase BT of CEO	61.4	12.5	12.2	6.2	7.8
Increase BT of administrative staff	79.5	9.8	6.0	2.4	2.3
Increase BT of sales personnel	71.1	9.9	8.6	6.5	3.9
Increase BT of others (specialists)	75.0	10.5	7.5	4.2	2.9

Vignoles 2005), health status (Riphahn et al. 2003), transportation (Hensher et al. 2010) and mobilisation time during a hurricane (Sadri et al. 2013). To the best of our knowledge, this work is the first attempt to apply this type of model to a potential increase in business travel due to a high-speed train project.

Following the work presented in Greene, Hensher (2009), we consider a latent variable y^* that captures how AlpTransit will modify firms' business relationships. This phenomenon can be described by the following latent regression model:

$$y_i^* = \beta' x_i + \epsilon_i$$

and is observed in discrete form through a censoring mechanism. In particular, our latent variable is represented by a discrete and ordinal indicator y_i:

$$y_i = j \quad \text{if} \quad \mu_{j-1} < y_i^* < \mu_j \qquad \text{for } j = 1, \ldots, 5$$

y_i is the self-assessed likelihood on a 5 point Likert scale of increasing BT for each employee category due to AlpTransit.

The sample observations (firms) are labelled $i = 1, \ldots, n$; the vector x_i contains all our explanatory variables, which are assumed to be strictly exogenous of ϵ_i and are described in Table 3. The vector of unknown parameters β and the thresholds μ_j (with $j = 1, \ldots, 5$) are the object of estimation and inference.

Following the literature, we include among the regressors the number (measured on a Likert scale from 1 to 5) and location of suppliers and customers[3], firms' characteristics: age, size, sector (in particular, manufacturing, business services, other activities in secondary and in tertiary sectors) and spatial organisation of the firm (if the firm is a branch, sole proprietorship or headquarters). In addition, we introduce frequency and destination (Milan, Ticino and Zurich) of current business travel as explanatory variables[4]. Some characteristics of current BT are taken into account such as the overnight stay in Zurich (due to the reduced travel time, firms could substitute them with more one-day trips) and the current use of train or car to go to the north of the Alps. 74% of respondents declared currently using car more than 50% of the times to go to Zurich, while 33% of companies answered using the train more than half of times. Only 1% of firms indicated flight as means of transport to go to Zurich, but only occasionally. Finally, we include a variable indicating the distance in kilometres from the nearest AlpTransit station in Ticino (Lugano or Bellinzona) to take into account another element of access to HST.

We assume that the error term ϵ_i is IID Logistic distributed with mean 0, scale parameter 1 and cumulative distribution function $\Lambda(\epsilon_i|x_i) = \Lambda(\epsilon_i)$.

[3] We do not introduce in the estimated models the variables "Suppliers in Ticino", "Suppliers in Italy", "Suppliers and customers in Western Europe", since only very few firms answered to these questions, resulting in coefficients that are not statistically significant.

[4] The variable "CEO current BT to Milan" is not introduced in the estimated model, since it is highly correlated with the variable "Sales personnel current BT to Milan".

Table 3: Explanatory variables

Variables	Operationalization	Mean	Std. Dev.
Suppliers in Ticino	How many suppliers firm has in Ticino	1.47	1.05
Suppliers in Italy	How many suppliers firm has in Italy	1.33	0.79
Suppliers in Zurich	How many suppliers firm has in Zurich	1.208	0.60
Suppliers in East Europe	How many suppliers firm has in East Europe	1.057	0.34
Suppliers in Western Europe	How many suppliers firm has in W. Europe	1.23	0.66
Clients in Ticino	How many clients firm has in Ticino	3.63	1.35
Clients in Zurich	How many clients firm has in Zurich	1.64	0.94
Clients in Italy	How many clients firm has in Italy	1.82	1.01
Clients in East Europe	How many clients firm has in East Europe	1.27	0.62
Clients in Western Europe	How many clients firm has in W. Europe	1.59	0.95
Services	Firm belongs to business services (financial and insurance activities, real estate, administrative, scientific and professional activities) – DV	0.258	0.44
Manufacturing	Firm belongs to manufacturing sector – DV	0.124	0.33
Other tertiary	Firm belongs to tertiary sector, business services excluded – DV	0.464	0.50
Other secondary	Firm belongs to secondary sector, manufacturing excluded – DV	0.153	0.36
Size	How many workers the plant in Ticino has	16.37	55.37
Age	How old the firm is	21	26.54
Sole proprietorship	Firm is a sole proprietorship – DV	0.78	0.41
Branch	Firm is a branch of a company group – DV	0.12	0.32
Headquarters	Firm is a headquarters of a company group – DV	0.10	0.30
Other plant in Zurich	Firm has another plant in Zurich – DV	0.048	0.22
CEO current BT to Ticino	Current BT frequency of CEO in Ticino	4.23	1.82
CEO current BT to Zurich	Current BT frequency of CEO to Zurich	2.06	1.19
CEO current BT to Milan	Current BT frequency of CEO to Milan	2.13	1.39
Admin. current BT to Ticino	Current BT frequency of admin. staff in Ticino	2.59	1.86
Admin. current BT to Zurich	Current BT frequency of admin. staff to Zurich	1.32	0.74
Admin. current BT to Milan	Current BT frequency of admin. staff to Milan	1.29	0.78
Comm. current BT to Ticino	Current BT frequency of sales personnel in Ticino	3.16	2.03
Comm. current BT to Zurich	Current BT frequency of sales personnel to Zurich	1.69	1.14
Comm. current BT to Milan	Current BT frequency of sales personnel to Milan	1.71	1.25
Other current BT to Ticino	Current BT frequency of specialists in Ticino	2.85	2.08
Other current BT to Zurich	Current BT frequency of specialists to Zurich	1.45	0.91
Other current BT to Milan	Current BT frequency of specialists to Milan	1.43	0.96
Current use of train to Zurich	Current use of train to go to Zurich (in percentage)	0.14	0.29
Current use of car to Zurich	Current use of car to go to Zurich (in percentage)	0.29	0.41
Overnight stay in Zurich	Current overnight stay in Zurich after a business meeting	0.14	0.35
Distance	Distance in km from Alptransit station	11.9	11.45

Notes: DV ... dummy variable

Given these assumptions, the probabilities associated with the observed outcomes are:

$$\begin{aligned} \text{Prob}[y_i = j|x_i] &= \text{Prob}[\mu_{j-1} < y_i^* < \mu_j] = \text{Prob}[\mu_{j-1} < \beta^{'} x_i + \epsilon_i < \mu_j] \\ &= \text{Prob}[\epsilon_i < \mu_j - \beta^{'} x_i] - \text{Prob}[\epsilon_i < \mu_{j-1} - \beta^{'} x_i] \\ &= \Lambda[\mu_j - \beta^{'} x_i] - \Lambda[\mu_{j-1} - \beta^{'} x_i] \end{aligned}$$

with $j = 1, \ldots, 5$. For identification purposes, we impose that $\mu_{(j-1)} < \mu_j$; $\mu_0 = -\infty$ and $\mu_5 = +\infty$.

The log-likelihood function, based on the previous implied probabilities, is:

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} \log[\Lambda(\mu_j - \beta' x_i) - \Lambda(\mu_{j-1} - \beta' x_i)]$$

where $m_{ij} = 1$ if $y_i = j$ and 0 otherwise. Using the maximum likelihood estimator (MLE), it is possible to estimate the parameters β and μ.

Table 4: Estimation results of the empirical models

Variables	Ordered logit			
	CEO	Admin.	Comm.	Others
Suppliers in Zurich	-0.203 (0.18)	-0.043 (0.21)	-0.202 (0.19)	-0.071 (0.19)
Suppliers in East Europe	-0.44 (0.33)	-1.313** (0.57)	-0.652* (0.39)	-0.62 (0.39)
Clients in Ticino	0.117 (0.08)	0.07 (0.10)	0.054 (0.09)	0.063 (0.09)
Clients in Zurich	0.28*** (0.1)	0.207* (0.11)	0.325*** (0.11)	0.151 (0.11)
Clients in Italy	-0.254** (0.11)	-0.093 (0.13)	-0.219* (0.11)	-0.107 (0.12)
Clients in East Europe	0.031 (0.14)	0.321** (0.16)	-0.003 (0.16)	0.06 (0.17)
Manufacturing	fixed	fixed	fixed	fixed
Services	1.123*** (0.35)	0.726* (0.44)	0.732** (0.38)	0.843** (0.39)
Other secondary	1.08*** (0.35)	1.081** (0.43)	1.03*** (0.37)	0.872** (0.38)
Other tertiary	0.463 (0.33)	0.297 (0.41)	0.543 (0.35)	0.346 (0.38)
Size	-0.0006 (0.0013)	-0.002 (0.001)	-0.0023 (0.002)	-0.0006 (0.0014)
Age	-0.0015 (0.0033)	-0.003 (0.004)	0.005 (0.003)	0.006* (0.003)
Headquarters	fixed	fixed	fixed	fixed
Sole proprietorship	-0.013 (0.28)	0.297 (0.36)	0.065 (0.30)	-0.078 (0.32)
Branch	0.10 (0.39)	0.868* (0.47)	0.233 (0.41)	0.048 (0.43)
Other plant in Zurich	-0.224 (0.47)	0.09 (0.51)	-0.22 (0.47)	-0.106 (0.5)
CEO current BT to Ticino	0.173*** (0.06)	-0.126* (0.08)	0.051 (0.07)	-0.0009 (0.07)
CEO current BT to Zurich	0.37*** (0.11)	0.33*** (0.13)	0.094 (0.12)	0.138 (0.13)
Admin. current BT to Ticino	-0.001 (0.06)	0.235*** (0.07)	0.104 (0.06)	0.085 (0.07)
Admin. current BT to Zurich	0.24* (0.14)	0.47*** (0.16)	0.09 (0.15)	0.33** (0.16)
Admin. current BT to Milan	-0.27* (0.14)	-0.013 (0.16)	-0.398** (0.14)	-0.573*** (0.16)
Comm. current BT to Ticino	0.03 (0.06)	0.052 (0.08)	0.114* (0.07)	-0.013 (0.07)
Comm. current BT to Zurich	-0.045 (0.11)	-0.0012 (0.14)	0.211* (0.13)	-0.079 (0.14)
Comm. current BT to Milan	-0.12 (0.10)	-0.14 (0.13)	0.391*** (0.11)	0.101 (0.12)
Other current BT to Ticino	-0.099* (0.06)	-0.048 (0.07)	-0.037 (0.06)	0.088 (0.07)
Other current BT to Zurich	-0.085 (0.13)	0.045 (0.15)	0.071 (0.13)	0.268** (0.14)
Other current BT to Milan	0.345*** (0.12)	0.092 (0.14)	0.206* (0.12)	0.479*** (0.13)
Current use of car to Zurich	0.51** (0.25)	0.609** (0.31)	0.349 (0.27)	0.085 (0.29)
Current use of train to Zurich	1.67*** (0.31)	0.967*** (0.36)	1.13*** (0.32)	1.019*** (0.34)
Overnight stay in Zurich	0.081 (0.23)	0.143 (0.28)	0.472* (0.25)	0.49* (0.27)
Distance km from	0.003 (0.007)	0.004 (0.009)	0.007 (0.008)	0.003 (0.008)
Alptransit station	0.003 (0.007)	0.004 (0.009)	0.007 (0.008)	0.003 (0.008)
Constant	-	-	-	-
Observations	696	696	696	696
Final Log-likelihood	-722.68	-448.901	-595.329	-538.816
R^2/Adjusted ρ^2	0.35	0.51	0.42	0.44

Notes: Std. errors in parenthesis $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

6 Results and discussion

Table 4 presents the results of the ordered logit models for the four employee categories. One of the main results of the analysis concerns the spatial structure of firms' current business travel. In particular, the greater a firm's BT within Ticino or a firm's BT to Zurich for one employee category currently are, the higher is the intention to travel more for the same category. We can call this "direct effect": AlpTransit will consolidate the travel behaviour of all these categories. It is interesting to notice that BT in Ticino and to Zurich have both positive and significant coefficients, if we look at the direct effects in all categories. This suggests that the two economies (Ticino and Zurich) are linked: not only will having current frequent business relationships with Zurich increase future travel, but also current trips within Ticino will boost such relationships.

We can also identify an "indirect effect", i.e. how current BT of one category influences the future BT of other categories. In particular, results show that the more frequently a CEO currently travels to Zurich, the higher the probability of more administrative staff also travelling. In addition, the more frequently administrative staff currently travel to Zurich, the higher the intention of increasing CEO and specialists' BT.

These "indirect effects" could be explained by the fact that different categories of employees travel for different purposes (Aguilera 2008): for example, after CEOs have

Table 5: Direct and indirect effects. Marginal effects for $y = 5$

Variables	CEO	Admin.	Comm.	Others
CEO current BT to Zurich	0.015***	0.0032**	0.0017	0.002
Admin. current BT to Zurich	0.009*	0.0045***	0.0016	0.0049*
Comm. current BT to Zurich	-0.002	-0.00001	0.0037*	-0.001
Other current BT to Zurich	-0.004	0.00043	0.0007	0.0039*

Notes: ***$p < 0.01$, **$p < 0.05$, *$p < 0.10$

established a strategic contact or have signed a new contract, they may leave the concrete tasks to the operational arm (administrative staff), increasing BT of that category.

Table 5 presents the marginal effects for the highest value of the ordinal dependent variable ($y = 5$), highlighting the direct and indirect effects.

Concerning the importance of a specific sector, being a business service enterprise (financial and insurance activities, real estate, administrative, scientific and professional activities), i.e. activities which support companies, positively influences (if compared to manufacturing) the intention to travel more with AlpTransit for all categories. This is in line with the literature; in particular, the greatest impact of HST is likely if the service sector is already a key economic characteristic of the region (Harman 2006), as is the case for the service sector in Ticino.

Moreover, the variable "other activities in the secondary sector" (mainly construction) is significant for all the models and has a positive sign: construction activities show a higher propensity toward increasing travel with AlpTransit (according to our data some professional profiles, like plant designers and tile layers, travel a lot).

Looking at the variables which indicate the current transport mode chosen to go to Zurich, both use of train and use of car positively affect the probability of travelling more with AlpTransit in the future. For the variable "current use of car to Zurich", the coefficient is not statistically significant for the "sales personnel" and "specialists" model. Therefore, not only those using the train at present, but also CEOs and administrative staff who are using cars, will travel more to Zurich via high-speed train.

In order to interpret the magnitude of the estimated coefficients for their effect on the categories of the ordinal dependent variable (i.e. on the propensity to increase BT), the marginal effects need to be computed. Table 6 presents a selection of the marginal effects for the last two ordinal categories of the dependent variable for the four employee groups. In general, the marginal effects are larger for the CEO model than for the other models. In particular, the likelihood of increasing CEO business travel due to AlpTransit (considering $y = 4$ and $y = 5$ together), increases by 12% on average for firms belonging to business services, all else being equal; while for the other employee categories, the same percentage is 2% for administrative staff, 4.8% for sales personnel and 4% for specialists. Concerning the respondent's current transport habits, an increase of 1% in the current use of train by CEOs, is associated with a 14% growth in the likelihood of increasing CEO business travel due to AlpTransit. Instead, an increase of 1% in the current use of train for non-CEOs, is associated with a lower growth in the likelihood of using AlpTransit. Specifically, a 2% for administrative staff, 6.4% for sales personnel and 4% for specialists.

Other interesting results are found using variables related to the geographical distribution of clients and suppliers (Table 4). What is clearly observable is that a higher number of current clients in Zurich correlates to a higher intention to travel more to Zurich with AlpTransit (the coefficients are positive and significant for three out of four models). Therefore, AlpTransit will consolidate business relations between Ticino firms and Zurich (the major Swiss economic pole), as expected. Indeed, close communication between business partners, which often implies face-to-face contacts, turns out to be essential for successful business transactions (Cristea 2011). The results from the suppliers require caution in the interpretation, since only a limited number of firms answered to these questions.

Being a branch positively affects the probability to travel more because of AlpTransit if compared with the reference category (being the headquarters), but only for the

Table 6: Marginal effects (selected results)

Variables	CEO		Admin.	
	y = 4	y = 5	y = 4	y = 5
Clients in Zurich	0.011***	0.011***	0.0025*	0.002*
Services	0.056***	0.061**	0.011*	0.0078*
Other secondary	0.057***	0.064**	0.019*	0.016*
Current use of car to Zurich	0.021**	0.021**	0.0074*	0.0058*
Current use of train to Zurich	0.069***	0.069***	0.012**	0.0093**

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

Variables	Comm.		Others	
	y = 4	y = 5	y = 4	y = 5
Clients in Zurich	0.013***	0.0057***	0.0038	0.0022
Services	0.033*	0.015*	0.026*	0.0156*
Other secondary	0.054**	0.026*	0.029*	0.0178*
Current use of car to Zurich	0.013	0.006	0.0021	0.0012
Current use of train to Zurich	0.044***	0.02***	0.026***	0.0151***

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$

administrative staff model. This result could be explained by the purpose of travel, which is mainly internal training for branches.

The variable indicating the presence of other plants of the same company in Zurich is not significant, but this is probably due to the low number of firms with this characteristic among our respondents.

The overnight stay variable is statistically significant only for the last two columns in Table 4 and is positive: currently, after a meeting, employees stay overnight in Zurich, thus positively influencing the intention of increasing travel for sales personnel and specialists, if compared with those firms that currently do not remain overnight. We can imagine for sales personnel and specialists that are less "eminent" categories in a firm, there is a sort of substitution effect: they will no longer stay overnight in Zurich, which is quite expensive, but they will have one-day trips more frequently.

Finally, in order to support our hypothesis that professional status and hierarchical position influence business travel, we estimate a Pooled OLS model with cluster standard errors. Results are shown in Appendix A. The dummy variable for CEO is taken as a reference and the variables for the other three employee categories are introduced into the model. The estimated coefficients of administrative staff, sales personnel and others (specialists) are all statistically significant and negative, meaning that there is a difference between these categories and the CEO in the propensity to increase business travel after AlpTransit. Moreover, if we test the equivalence among these three coefficients, we can reject the null hypothesis that they are equal.

7 Conclusions

As demonstrated by the literature review, accessibility improvements have very complex but measurable implications on regional development. However, while these effects are empirically demonstrated on an aggregate level, implications of large infrastructure projects for individual firms have received little attention. In this research, we demonstrate that in the case of a clearly defined and predictable change in services of a new infrastructure project, it is possible to identify future adaptation of business travel to improved accessibility.

The main goal of this paper is to understand how the propensity of employees in various types of firms to travel between a small city and a dominant one will change after AlpTransit, the high-speed railway project in Switzerland. Using new micro data, we

apply ordered logit models in order to investigate this issue. We obtain differentiated results among employees with different functions in various types of firms. Being a business services firm positively affects the probability of increasing face-to-face contacts. This result is in line with the literature: firms providing support to other companies take more advantage of high-speed trains (Blum et al. 1997, Jones 2007) and exploit the opportunities linked to an enlarged market (see the experience of Lyon companies: Harman 2006). Moreover, in this specific case, Ticino and Zurich have a common specialisation in financial, professional and scientific activities: both the localisation ratio and the specialisation ratio are quite high in both regions for those industries. We conclude that the existence of common clusters encourages the development of business relations between the two territories.

Another interesting finding from the study is that a firm's current level of business travel between Ticino and Zurich as well as intensive travel activity within Ticino, encourages future face-to-face contacts. We conclude that, in line with our theoretical considerations, BT habits influence future ones, reinforced by interaction effects among higher- and lower-level employee categories.

Overall, our results are in line with those in the ex-post evaluation literature. Therefore, we can conclude that our experiment is realistic and the ex-ante evaluation is credible. This might be due to the fact that the project considered, AlpTransit, intervenes in a context of existing strong business ties along the link, and that the future services provided in terms of time-table, frequency of service and prices were known in advance.

There are limits to the analysis of foreseen prospective impacts. Future research after the opening of AlpTransit will have to verify our findings for Ticino and include an equivalent approach to the analysis of the behaviour of firms in Zurich. After all, given the 'double importance' of accessibility, Ticino will also be closer for Zurich firms.

In conclusion, an ex-ante evaluation of the effects of new infrastructure is possible as long as the focus is on a specific travel category (in our case, business travel).

Overall, our study foresees positive impacts on future business travels due to HST, differentiating by functions of employees and current travel intensity. We therefore expect positive effects of the opening of the new connection on the Ticino economy.

References

Aarts H, Dijksterhuis A (2000) The automatic activation of goal-directed behaviour: The case of travel habit. *Journal of Environmental Psychology* 20: 75–82. CrossRef.

Aarts H, Verplanken B, van Knippenberg A (1998) Predicting behavior from actions in the past: Repeated decisions making or a matter of habit? *Journal of Applied Social Psychology* 28: 1355–1374. CrossRef.

Aguilera A (2008) Business travel and mobile workers. *Transportation Research Part A* 42: 1109–1116. CrossRef.

Axhausen KW, König A, Abay G, Bates JJ, Bierlaire M (2006) Swiss value of travel time savings. ETH Research Collection, Zürich

Banister D, Thurstain-Goodwin M (2011) Quantification of the non-transport benefits resulting from rail investment. *Journal of Transport Geography* 19[2]: 212–223. CrossRef.

Beaverstock JV, Derudder B, Faulconbridge JR, Witlox F (2009) International business travel: Some explorations. *Geografiska Annaler Series B* 91[3]: 193–202. CrossRef.

Blum U, Haynes KE, Karlsson C (1997) The regional and urban effects of high-speed trains. *The Annals of Regional Science* 31: 1–20. CrossRef.

Bruinsma F, Rietveld P (1998) The accessibility of cities in European infrastructure network. In: Bruinsma F, Rietveld P (eds), *Is transport infrastructure effective?* Springer, Heidelberg. CrossRef.

Carbo JM, Graham DJ, Casas AD, Melo PC (2019) Evaluating the causal economic impacts of transport investments: Evidence from the Madrid-Barcelona high speed rail corridor. *Journal of Applied Statistics* 46[9]: 1714–1723. CrossRef.

Choo S, Lee TY, Mokhtarian PL (2007) Do transportation and communications tend to be substitutes, complements, or neither? U.S. consumer expenditures perspective, 1984-2002. *Transportation Research Record* 2010: 123–132

Cristea AD (2011) Buyer-seller relationships in international trade: Evidence from US states' exports and business-class travel. *Journal of International Economics* 84: 207–220. CrossRef.

De Bok M, Sanders F (2005) Firm relocation and accessibility of locations. Empirical results from the Netherlands. *Journal of Transportation Research Board* 1902: 35–43

Greene W, Hensher DA (2009) *Modeling Ordered Choice*. Cambridge University Press, Cambridge. CrossRef.

Gustafson P (2012) Managing business travel: Developments and dilemmas in corporate travel management. *Tourism Management* 33: 276–284. CrossRef.

Gutiérrez J (2001) Location, economic potential and daily accessibility: An analysis of the accessibility impact of the high-speed line Madrid-Barcelona-French border. *Journal of Transport Geography* 9: 229–242. CrossRef.

Harman R (2006) High-speed trains and the development and regeneration of cities. Greengauge21, London. http://www.greengauge21.net/wp-content/uploads/hsr-regneration-of-cities.pdf

HBR – Harvard Business Review Analytic Services (2009) Managing across distance in today's economic climate: The value of face-to-face communication. Harvard Business School Publishing, Harvard University

Hensher DA, Mulley C, Yahya N (2010) Passenger experience with quality-enhanced bus service: The tyne and wear superoute services. *Transportation* 37: 239–256. CrossRef.

Hess S, Daly A (2014) *Handbook of choice modelling*. Edward Elgar, Cheltenham. CrossRef.

Hovhannisyan N, Keller W (2015) International business travel: An engine of innovation? *Journal of Economic Growth* 20[1]: 75–104. CrossRef.

Hugoson P (2001) Interregional business travel and the economics of business interaction. Jönköping International Business School, Jönköping. http://hj.diva-portal.org/smash/get/diva2:3898/FULLTEXT01.pdf

Jones A (2007) More than 'managing across borders?' the complex role of face-to-face interaction in globalizing law firms. *Journal of Economic Geography* 7[3]: 223–246. CrossRef.

Kawamura K (2004) Transportation needs, location choice and perceived accessibility for businesses. *Journal of Transportation Research Board* 1898: 202–210

Kobayashi K, Okumura M (1997) The growth of city systems with high-speed railway systems. *The Annals of Regional Science* 31: 39–56. CrossRef.

Leitham S, McQuaid RW, Nelson JD (2000) The influence of transport on industrial location choice: A stated preference experiment. *Transportation Research Part A* 34: 515–535. CrossRef.

Machikita T, Ueki Y (2010) The impacts of face-to-face and frequent interactions on innovation: Evidence from upstream-downstream relations. *International Journal of Institutions and Economies* 3[3]: 519–548

Machin S, Vignoles A (2005) *What's the good of education? The economics of education in the UK*. Princeton University Press, Princeton

Maggi R (1989) Towards an economic theory of barriers to communication. *Papers of the Regional Science Association* 66: 131–141. CrossRef.

Marti M, Sommer H, Maggi R (2007) Erreichbarkeit und regionalwirtschaftliche entwicklung. Jahrbuch 2007 Schweizerische Verkehrswirtschaft. St. Gallen, 221-233

Mazzeo G (2012) Impact of high speed trains on the hierarchy of European cities. *Jahrbuch für Regionalwissenschaft* 32: 159–173. CrossRef.

McCann P (2001) *Modern urban and regional economics*. Oxford University Press, Oxford

Poole J (2010) Business travel as an input to international trade. Mimeo, University of California Santa Cruz

Riphahn R, Wambach A, Million A (2003) Incentive effects on the demand for health care: A bivariate panel count data estimation. *Journal of Applied Economics* 18[4]: 387–405. CrossRef.

Sadri AM, Ukkusuri SV, Murray-Tuite P (2013) A random parameter ordered probit model to understand the mobilization time during hurricane evacuation. *Transportation Research Part C* 32: 21–30. CrossRef.

Swarbrooke J, Horner S (2001) *Business Travel and Tourism*. Butterworth-Heinemann, Oxford. CrossRef.

Targa F, Clifton KJ, Mahmassani HS (2006) Influence of transportation access on individual firm location decisions. *Journal of the Transportation Research Board* 1977: 179–189. CrossRef.

Törnqvist G (1984) Contact potentials in the european system of cities. Contribution to Metropolitan Study, 16 CP-84-55

UNWTO – United Nations World Tourism Organization (2012) Tourism highlights. UNWTO publication, Madrid

Ureña JM, Menerault P, Garmendia M (2009) The high-speed rail challenge for big intermediate cities: A national, regional and local perspective. *Cities* 26: 266–279. CrossRef.

Vickerman R, Spiekermann K, Wegener M (1999) Accessibility in economic development in Europe. *Regional Studies* 33[1]: 1–15. CrossRef.

Vickerman R, Ulied A (2009) Indirect and wider economic impacts of high-speed rail. Economia y sociedad, Fundacion BBVA, Valencia

Wardman M, Batley R, Laird J, Mackie P, Fowkes T, Lyons G, Bates J, Eliasson J (2013) Valuation of travel time savings for business travellers. (report for the UK Department for Transport). Institute for Transport Studies, University of Leeds, Leeds. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/-attachment_data/file/251998/annexes-for-main-report-dft-005.pdf

Welch DE, Worm W (2005) International business travellers: a challenge for IHRM. In: Stahl GK, Björkman I (eds), *Handbook of research in human resource management*. Edward Elgar, Cheltenham, UK, 283–301. CrossRef.

Willingers J, Floor H, van Wee B (2007) Accessibility indicators for location choices of offices: An application to the intraregional distributive effects of high-speed rail in the Netherlands. *Environment and Planning A* 39: 2086–2098. CrossRef.

WTTC – World, Travel and Tourism Council (2011) Business travel: A catalyst for economic performance. London. https://www.deplacementspros.com/attachment/281391/

A Appendix: Estimation results of Pooled OLS model

Variables	Pooled OLS
CEO	fixed
Administrative staff	-0.481*** (0.03)
Sales personnel	-0.244*** (0.033)
Others (specialists)	-0.37*** (0.037)
Suppliers in Zurich	-0.055 (0.06)
Suppliers in East Europe	-0.135** (0.06)
Clients in Ticino	0.023 (0.03)
Clients in Zurich	0.094** (0.05)
Clients in Italy	-0.071* (0.04)
Clients in East Europe	0.013 (0.06)
Manufacturing	fixed
Services	0.438*** (0.11)
Other secondary	0.363*** (0.103)
Other tertiary	0.245*** (0.097)
Size	-0.0011** (0.0005)
Age	0.0001 (0.0012)
Headquarters	fixed
Sole proprietorship	0.057 (0.10)
Branch	0.244* (0.14)
Other plant in Zurich	-0.113 (0.19)
CEO current BT to Ticino	0.032 (0.02)
CEO current BT to Zurich	0.085 (0.06)
Admin. current BT to Ticino	0.026 (0.02)
Admin. current BT to Zurich	0.141* (0.08)
Admin. current BT to Milan	-0.129** (0.05)
Comm. current BT to Ticino	0.014 (0.02)
Comm. current BT to Zurich	0.047 (0.05)
Comm. current BT to Milan	0.054 (0.03)
Other current BT to Ticino	-0.015 (0.02)
Other current BT to Zurich	0.03 (0.06)
Other current BT to Milan	0.129*** (0.05)
Current use of car to Zurich	0.08 (0.11)
Current use of train to Zurich	0.581*** (0.15)
Overnight stay in Zurich	0.191 (0.12)
Distance km from Alptransit station	0.002 (0.003)
Constant	0.651** (0.26)
Observations	2784
R-squared	0.22

Std. errors in parenthesis $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

REGION
The Journal of ERSA
Powered by WU

Volume 6, Number 3, 2019, 17–37
DOI: 10.18335/region.v6i3.276

journal homepage: region.ersa.org

Demonstrating the utility of machine learning innovations in address matching to spatial socio-economic applications*

Sam Comber[1]

[1] University of Liverpool, Liverpool, United Kingdom

Received: 16 August 2019/Accepted: 30 December 2019

Abstract. The last decade has heralded an unprecedented rise in the number, frequency and availability of data sources. Yet they are often incomplete, meaning data fusion is required to enhance their quality and scope. In the context of spatial analysis, address matching is critical to enhancing household socio-economic and demographic characteristics. Matching administrative, commercial, or lifestyle data sources to items such as household surveys has the potential benefits of improving data quality, enabling spatial data visualisation, and the lowering of respondent burden in household surveys. Typically when a practitioner has high quality data, unique identifiers are used to facilitate a direct linkage between household addresses. However, real-world databases are often absent of unique identifiers to enable a one-to-one match. Moreover, irregularities between the text representations of potential matches mean extensive cleaning of the data is often required as a pre-processing step. For this reason, practitioners have traditionally relied on two linkage techniques for facilitating matches between the text representations of addresses that are broadly divided into deterministic or mathematical approaches. Deterministic matching consists of constructing hand-crafted rules that classify address matches and non-matches based on specialist domain knowledge, while mathematical approaches have increasingly adopted machine learning techniques for resolving pairs of addresses to a match. In this notebook we demonstrate methods of the latter by demonstrating the utility of machine learning approaches to the address matching work flow. To achieve this, we construct a predictive model that resolves matches between two small datasets of restaurant addresses in the US. While the problem case may seem trivial, the intention of the notebook is to demonstrate an approach that is reproducible and extensible to larger data challenges. Thus, in the present notebook, we document an end-to-end pipeline that is replicable and instructive towards assisting future address matching problem cases faced by the regional scientist.

1 Introduction

Our overarching objective is to demonstrate how machine learning can be integrated into the address matching work flow. By definition, address matching pertains to the process of resolving pairs of records with a spatial footprint. While geospatial matching links the geometric representations of spatial objects, address matching typically involves linking the text-based representations of address pairs. The utility of address matching, and

*This paper is available as computational notebook on the REGION webpage.

record linkage in general, lies in the ability to unlock attributes from sources of data that cannot be linked by traditional means. This is often because the datasets lack a common key to resolve a join between the address of a premise. Two example applications of address matching uses include: the linkage of historical censuses across time for exploring economic and geographic mobility across multiple generations (Ruggles et al. 2018), and exploring how early-life hazardous environmental exposure, socio-economic conditions, or natural disasters impact the health and economic outcomes of individuals living in particular residential locations (Cayo, Talbot 2003, Reynolds et al. 2003, Baldovin et al. 2015).

For demonstrative purposes, we rely on small a set of addresses from the Fodors and Zagat restaurant guides that contain 112 matched addresses for training a predictive model that resolves address pairs to matches and non-matches. In a real-world application, training a machine learning model on a small sample of matched addresses could be used to resolve matches between the remaining addresses of a larger dataset. While we use the example of restaurant addresses, these could easily be replaced by addresses from a far less trivial source and the work flow required to implement the address matching exercise would remain the same. Therefore, it is the intention of this guide to provide insight on how the work flow of a supervised address matching work flow proceeds, and to inspire interested users to scale the supplied code to larger and more interesting problems.

2 Packages and dependencies

[1]:
```
%matplotlib inline
import os
import uuid
import warnings
from IPython.display import HTML

# load external libraries
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import jellyfish
import recordlinkage as rl
import seaborn as sns
from postal.parser import parse_address # CRF parser
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_validate, train_test_split
from sklearn.metrics import precision_score, recall_score, f1_score, confusion_matrix

# configure some settings
np.random.seed(123)
sns.set_style('whitegrid')
pd.set_option('display.max_colwidth', -1)
warnings.simplefilter(action='ignore', category=FutureWarning)

def hover(hover_color="#add8e6"):
    return dict(selector="tbody tr:hover",
            props=[("background-color", "%s" % hover_color)])

# table CSS
styles = [
    #table properties
    dict(selector=" ",
         props=[("margin","0"),
                ("font-family",'"Helvetica", "Arial", sans-serif'),
                ("border-collapse", "collapse"),
                ("border","none"), ("border-style", "hidden")]),
    dict(selector="td", props = [("border-style", "hidden"),
                                 ("border-collapse", "collapse")]),

    #header color
    dict(selector="thead",
         props=[("background-color","#a4dbc8")]),
```

```
    #background shading
    dict(selector="tbody tr:nth-child(even)",
         props=[("background-color", "#fff")]),
    dict(selector="tbody tr:nth-child(odd)",
         props=[("background-color", "#eee")]),

    #header cell properties
    dict(selector="th",
         props=[("text-align", "center"),
                ("border-style", "hidden"),
                ("border-collapse", "collapse")]),

    hover()
]
```

3 Data loading, cleaning and segmentation

To begin our exercise we specify the file location that contains the entirety of the 112 Zagat and Fodor matched address pairs. This file can be downloaded from the dedicated Github repository that accompanies the paper (https://github.com/SamComber/address_matching_workflow) using the `wget` command.

[2]:
```
! wget https://raw.githubusercontent.com/SamComber/address_matching_workflow/master/
...zagat_fodor_matched.txt
```

[2]:
```
--2019-12-21 09:11:31--  https://raw.githubusercontent.com/SamComber/address_matching_
workflow/master/zagat_fodor_matched.txt
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 199.232.56.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|199.232.56.133|:443
... connected.
HTTP request sent, awaiting response... 200 OK
Length: 19939 (19K) [text/plain]
Saving to: ‘zagat_fodor_matched.txt’

zagat_fodor_matched 100%[===================>]  19.47K  --.-KB/s    in 0.03s

2019-12-21 09:11:32 (670 KB/s) - ‘zagat_fodor_matched.txt’ saved [19939/19939]
```

[3]:
```
f = 'zagat_fodor_matched.txt'
```

Address matching is principally a data quality challenge. Similar to other areas of data analysis, when the quality of input data to the match classification is low, the output generated will typically be of low accuracy (Christen 2012). Problematically, most address databases we encounter in the real world are inconsistent, are missing of several values, and lack standardisation. Thus, a first step in the address matching work flow is to increase the quality of input data. In this way we increase the accuracy, completeness and consistency of our address records, which increases the ease in which they can be linked by the techniques we apply later on. Typically this stage begins by parsing the text representations of addresses into rows of a dataframe.

[4]:
```
# load matched addresses, remove comment lines and reshape into two columns
data = pd.read_csv(f, comment='#',
                   header=None,
                   names=['address']).values.reshape(-1, 2)

matched_address = pd.DataFrame(data, columns=['addr_zagat', 'addr_fodor'])
```

[5]:
```
print('{} matched addresses loaded.'.format(matched_address.shape[0]))
matched_address.head(10).style.set_table_styles(styles)
```

Table 1: Output produced by code 5

	addr_zagat	addr_fodor
0	Arnie Morton's of Chicago 435 S. La Cienega Blvd. Los Angeles 90048 310-246-1501 Steakhouses	Arnie Morton's of Chicago 435 S. La Cienega Blvd. Los Angeles 90048 310/246-1501 American
1	Art's Deli 12224 Ventura Blvd. Studio City 91604 818-762-1221 Delis	Art's Delicatessen 12224 Ventura Blvd. Studio City 91604 818/762-1221 American
2	Bel-Air Hotel 701 Stone Canyon Rd. Bel Air 90077 310-472-1211 Californian	Hotel Bel-Air 701 Stone Canyon Rd. Bel Air 90077 310/472-1211 Californian
3	Cafe Bizou 14016 Ventura Blvd. Sherman Oaks 91423 818-788-3536 French Bistro	Cafe Bizou 14016 Ventura Blvd. Sherman Oaks 91423 818/788-3536 French
4	Campanile 624 S. La Brea Ave. Los Angeles 90036 213-938-1447 Californian	Campanile 624 S. La Brea Ave. Los Angeles 90036 213/938-1447 American
5	Chinois on Main 2709 Main St. Santa Monica 90405 310-392-9025 Pacific New Wave	Chinois on Main 2709 Main St. Santa Monica 90405 310/392-9025 French
6	Citrus 6703 Melrose Ave. Los Angeles 90038 213-857-0034 Californian	Citrus 6703 Melrose Ave. Los Angeles 90038 213/857-0034 Californian
7	Fenix at the Argyle 8358 Sunset Blvd. W. Hollywood 90069 213-848-6677 French (New)	Fenix 8358 Sunset Blvd. West Hollywood 90069 213/848-6677 American
8	Granita 23725 W. Malibu Rd. Malibu 90265 310-456-0488 Californian	Granita 23725 W. Malibu Rd. Malibu 90265 310/456-0488 Californian
9	Grill The 9560 Dayton Way Beverly Hills 90210 310-276-0615 American (Traditional)	Grill on the Alley 9560 Dayton Way Los Angeles 90210 310/276-0615 American

[5]:
```
112 matched addresses loaded.

Output in Table 1
```

A series of data cleaning exercises will then modify the data in ways that support the application of the linkage techniques. This might involve writing data cleaning scripts that convert all letters to lowercase characters, delete leading and trailing whitespaces, remove unwanted characters and tokens such as punctuation, or using hard-coded look-up tables to find and replace particular tokens. All together coding these steps contributes towards a standard form between the two address databases the user is attempting to match. This is important because standards between the two sources of address data under consideration will typically differ due to different naming conventions.

In the following cell blocks, we execute these steps by standardising our addresses. More specifically, we remove non-address components, convert all text to lower case and remove punctuation and non-alphanumeric characters.

[6]:
```
# our rows contain non-address components such as phone number and
# restaurant type so lets parse these using regular expressions into new columns
zagat_pattern = r"(?P<address>.*?)(?P<phone_number>\b\d{3}\-\d{3}\-\d{4}\b)(?
...P<category>.*$)"
fodor_pattern = r"(?P<address>.*?)(?P<phone_number>\b\d{3}\/\d{3}\-\d{4}\b)(?
...P<category>.*$)"

matched_address[["addr_zagat", "phone_number_zagat", "category_zagat"]] =
...matched_address["addr_zagat"].str.extract(zagat_pattern)
matched_address[["addr_fodor", "phone_number_fodor", "category_fodor"]] =
...matched_address["addr_fodor"].str.extract(fodor_pattern)
```

[7]:
```
# standardise dataframe by converting all strings to lower case
matched_address = matched_address.applymap(lambda row : row.lower() if type(row) ==
... str else row)

# remove punctuation and non-alphanumeric characters
matched_address['addr_zagat'] = matched_address['addr_zagat'].str.replace('[^\w\s]','')
matched_address['addr_fodor'] = matched_address['addr_fodor'].str.replace('[^\w\s]','')
```

3.1 Segmentation of address string into field columns

After removing unwanted characters and tokens, our next step is to segment the entire address string into tagged attribute values. Addresses rarely come neatly formatted into sensible fields that identify each component, and so segmentation is a vital and often overlooked stage of the work flow. For example, an address might come in an unsegmented format such as "19 Water St. New York 11201". Our objective is then to segment (or label) this address into the appropriate columns for street number, street name, city and postcode. When we segment both sets of addresses from the datasets we intend to link, we build well-defined output fields that are suitable for matching.

In our case we use a statistical segmentation tool called **Libpostal** which is a Conditional Random Fields (CRFs) model trained on OpenStreetMap addresses. Before using the Python bindings, users are required to install the Libpostal C library first (see https://github.com/openvenues/pypostal#installation for installation instructions). CRFs are popular methods in natural language processing (NLP) for predicting sequence of labels across sequences of text inputs. Unlike discrete classifiers, CRFs model the probability of a transition between labels on "neighbouring" elements, meaning they take into account past and future address field states into the labelling of addresses into address fields. This mitigates a limitation of segmentation models such as hidden markov models (HMMs) called the *label bias problem*: "transitions leaving a given state to compete only against each other, rather than against all transitions in the model" (Lafferty et al. 2001). Take, for example, the business address for "1st for Toys, 244 Ponce de Leon Ave. Atlanta 30308". A naive segmentation model would incorrectly parse "1st" as a property number, whereas it actually completes the business name "1st for Toys", leading to an erroneous sequence of label predictions. When a CRFs has parsed "1st" and reaches the second token, "for", the model scores an $l \times l$ matrix where l is the maximal number of labels (or address fields) that can be assigned by the CRFs. In L, l_{ij} reflects the probability of the current word being labelled as i and the previous word labelled j (Diesner, Carley 2008). In a CRFs model, when the parser reaches the *actual* property number, "244", high scoring in the matrix indicates the current label should be a property number, and the previous label revised to a business name. For a more detailed account, see Comber, Arribas-Bel (2019).

To segment each address, we apply the `parse_address` function row-wise for both the Zagat and Fodors addresses. This generates a list of tuples (see below code block for an example of the first two addresses from the Zagat dataset) that we convert into dictionaries before finally reading these into a `pandas` dataframe.

[8]:
```
[[('arnie mortons of chicago', 'house'),
  ('435', 'house_number'),
  ('s la cienega blvd', 'road'),
  ('los angeles', 'city'),
  ('90048', 'postcode')],
 [('arts deli', 'house'),
  ('12224', 'house_number'),
  ('ventura blvd', 'road'),
  ('studio city', 'city'),
  ('91604', 'postcode')]]
```

[8]:
```
[[('arnie mortons of chicago', 'house'),
  ('435', 'house_number'),
  ('s la cienega blvd', 'road'),
  ('los angeles', 'city'),
  ('90048', 'postcode')],
 [('arts deli', 'house'),
  ('12224', 'house_number'),
  ('ventura blvd', 'road'),
  ('studio city', 'city'),
  ('91604', 'postcode')]]
```

[9]:
```
# parse address string using libpostal CRF segmentation tool
addr_zagat_parse = [parse_address(addr, country='us') for addr in
... matched_address.addr_zagat]
addr_fodor_parse = [parse_address(addr, country='us') for addr in
... matched_address.addr_fodor]

# convert to pandas dataframe
addr_zagat_parse = pd.DataFrame.from_records([{k: v for v, k in row} for row in
... addr_zagat_parse]).add_suffix('_zagat')
addr_fodor_parse = pd.DataFrame.from_records([{k: v for v, k in row} for row in
... addr_fodor_parse]).add_suffix('_fodor')

# vertical join of CRF-parsed addresses between both dataframes
matched_address = matched_address.join(addr_zagat_parse).join(addr_fodor_parse)
```

Given we know the match status of our training data, we can safely join the records back together once we have successfully segmented them. Moreover, as we know the match status in advance, we can assign unique IDs that we will use later to create a binary variable for indicating whether an address pair is matched or non-matched.

[10]:
```
# create unique ID for matched addresses, these will be used later to create a match
 status
uids = [str(uuid.uuid4()) for i in matched_address.iterrows()]

# the following two lines will assign the same uid to both columns, thus facilitating
 a match
addr_zagat_parse['uid'], addr_fodor_parse['uid'] = uids, uids
match_ids = pd.DataFrame({'zagat_id' : addr_fodor_parse['uid'], 'fodor_id' :
... addr_fodor_parse['uid']})
```

[11]:
```
# join match ids to main dataframe
matched_address = matched_address.join(match_ids)

# preview of our parsed dataframe with uids assigned
matched_address.head().style.set_table_styles(styles)
```

[11]:
```
Output in Table 2
```

4 Creation of candidate address pairs using a 'full index'

Once our addresses have met a particular standard of quality and are segmented into the desired address fields, the next step requires us to create candidate pairs of addresses that potentially resolve to the same address. In record linkage, this step is typically called indexing or blocking, and is required to reduce the number of address pairs that are compared. In doing so we remove pairs that are unlikely to resolve to true matches. To demonstrate the utility of blocking and why it is so important to address matching, we first create a **full index** which creates all possible combinations of address pairs. More concretely, a full index generates the Cartesian product between both sets of addresses. Conditional on the size of both dataframes, full blocking is highly computationally inefficient, and in our case we create $112 \times 112 = 12544$ candidate links; this has a complexity of $O(n^2)$. We demonstrate the full index method to motivate the desire for practitioners to implement more sophisticated blocking techniques.

4.1 Full index

Below, we instantiate an `Index` class before specifying the desired full index method for generating pairs of records. We then create the Cartesian join between the Zagat and Fodor addresses which creates a `MultiIndex` that links every Zagat address with every Fodor address.

Table 2: Output produced by code 11

Nr	addr_zagat	addr_fodor	phone_number_zagat	category_zagat	phone_number_fodor	category_fodor	city_zagat	city_district_zagat
0	arnie mortons of chicago 435 s la cienega blvd los angeles 90048	arnie mortons of chicago 435 s la cienega blvd los angeles 90048	310-246-1501	steakhouses	310/246-1501	american	los angeles	nan
1	arts deli 12224 ventura blvd studio city 91604	arts delicatessen 12224 ventura blvd studio city 91604	818-762-1221	delis	818/762-1221	american	studio city	nan
2	belair hotel 701 stone canyon rd bel air 90077	hotel belair 701 stone canyon rd bel air 90077	310-472-1211	californian	310/472-1211	californian	nan	nan
3	cafe bizou 14016 ventura blvd sherman oaks 91423	cafe bizou 14016 ventura blvd sherman oaks 91423	818-788-3536	french bistro	818/788-3536	french	sherman oaks	nan
4	campanile 624 s la brea ave los angeles 90036	campanile 624 s la brea ave los angeles 90036	213-938-1447	californian	213/938-1447	american	los angeles	nan

Continued (additional columns) on next page

Table 2 – *Continued from previous page*

Nr	house_zagat	house_number_zagat	postcode_zagat	road_zagat	suburb_zagat	city_fodor	city_district_fodor	house_fodor
0	arnie mortons of chicago	435	90048	s la cienega blvd	nan	los angeles	nan	arnie mortons of chicago
1	arts deli	12224	91604	ventura blvd	nan	studio city	nan	arts delicatessen
2	belair hotel	701	90077	stone canyon rd bel air	nan	nan	nan	hotel belair
3	cafe bizou	14016	91423	ventura blvd	nan	sherman oaks	nan	cafe bizou
4	campanile	624	90036	s la brea ave	nan	los angeles	nan	campanile

Nr	house_number_fodor	postcode_fodor	road_fodor	state_fodor	suburb_fodor
0	435	90048	s la cienega blvd	nan	nan
1	12224	91604	ventura blvd	nan	nan
2	701	90077	stone canyon rd bel air	nan	nan
3	14016	91423	ventura blvd	nan	nan
4	624	90036	s la brea ave	nan	nan

Nr	zagat_id	fodor_id
0	99bbcd03-ce45-40b5-907f-47f5ae16ae29	99bbcd03-ce45-40b5-907f-47f5ae16ae29
1	1b1e1ee1-c880-4722-abaa-4ec44c7d94a6	1b1e1ee1-c880-4722-abaa-4ec44c7d94a6
2	f2548f68-2326-4706-bdc1-dbfc265ecbf3	f2548f68-2326-4706-bdc1-dbfc265ecbf3
3	936687e2-1161-4ecd-98c3-b5ac620d8776	936687e2-1161-4ecd-98c3-b5ac620d8776
4	20a05006-d08e-4245-b014-ab2d0f552662	20a05006-d08e-4245-b014-ab2d0f552662

[12]:
```
indexer = rl.Index()
indexer.full()

# create cartesian join between zagat and fodor restaurant addresses
candidate_links = indexer.index(matched_address.city_zagat, matched_address.city_fodor)
```

[12]:
```
WARNING:recordlinkage:indexing - performance warning - A full index can result in large
number of record pairs.
```

[13]:
```
# this creates a two-level multiindex, so we name addresses from the zagat and fodor
 databases, respectively.
candidate_links.names = ['zagat', 'fodor']

print('{} candidate links created using full indexing.'.format(len(candidate_links)))
```

[13]:
```
12544 candidate links created using full indexing.
```

In practice, a full index creates a dataframe with 12,544 rows and thus creates candidate address pairs between every possible combination of address from both the Zagat and Fodor datasets. Once we generate this dataframe of potential matches, we create a match status column and assign a 1 to actual matched addresses and 0 to non-matches based on the unique IDs created earlier.

[14]:
```
# lets create a function we can reuse later on
def return_candidate_links_with_match_status(candidate_links):

    # we return a vector of label values for both the zagat and fodor restaurant
    IDs from the multiindex
    zagat_ids = candidate_links.get_level_values('zagat')
    fodor_ids = candidate_links.get_level_values('fodor')

    # now we create a new dataframe as long as the number of candidate links
    zagat = matched_address.loc[zagat_ids][['city_zagat', 'house_zagat',\
                                            'house_number_zagat', 'road_zagat',\
                                            'suburb_zagat', 'zagat_id']]
    fodor = matched_address.loc[fodor_ids][['city_fodor','house_fodor',\
                                            'house_number_fodor', 'road_fodor',\
                                            'suburb_fodor', 'fodor_id']]

    # vertically concateate addresses from both databases
    candidate_link_df = pd.concat([zagat.reset_index(drop=True),
    ... fodor.reset_index(drop=True)], axis=1)

    # next we create a match status column that we will use to train a machine
     learning model
    candidate_link_df['match_status'] = np.nan

    # assign 1 for matched, 0 non-matched
    candidate_link_df.loc[candidate_link_df['zagat_id'] ==
    ... candidate_link_df['fodor_id'], 'match_status'] = 1.
    candidate_link_df.loc[ ~(candidate_link_df['zagat_id'] ==
    ... candidate_link_df['fodor_id']), 'match_status'] = 0.

    return candidate_link_df

candidate_link_df = return_candidate_links_with_match_status(candidate_links)
```

4.2 Creation of comparison vectors from indexed addresses

To resolve addresses into matches and non-matches we generate comparison vectors between each candidate address pair. Each element of this comparison vector is a similarity metric used to assess the closeness of two address fields. In our case, we use **Jaro-Winkler similarity** because it has been observed to perform best on attributes containing named values (e.g., property names, street names, or city names) (Christen

2012, Yancey 2005). The Jaro similarity of two given address components a_1 and a_2 is given by

$$jaro\_sim = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|a_1|} + \frac{m}{|a_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

where $|a_i|$ is the length of the address component string a_i, m is the number of matching characters, and t is the number of transpositions required to match the two address components. We will create a function that makes use of the `jellyfish` implementation of Jaro-winkler similarity. Several other string similarity metrics are available and are optimised for particular use cases and data types. See Chapter 5 of Christen (2012) for an excellent overview.

[15]:
```
def jarowinkler_similarity(s1, s2):

    conc = pd.concat([s1, s2], axis=1, ignore_index=True)

    def jaro_winkler_apply(x):

        try:
            return jellyfish.jaro_winkler(x[0], x[1])
        # raise error if fields are empty
        except Exception as err:
            if pd.isnull(x[0]) or pd.isnull(x[1]):
                return np.nan
            else:
                raise err

    # apply row-wise to concatenated columns
    return conc.apply(jaro_winkler_apply, axis=1)
```

Before applying Jaro-Winkler similarity we need to choose columns that were segmented in **both** the Zagat and Fodor datasets.

[16]:
```
# lets take a look at the columns we have available
candidate_link_df.columns
```

[16]:
```
Index(['city_zagat', 'house_zagat', 'house_number_zagat', 'road_zagat',
       'suburb_zagat', 'zagat_id', 'city_fodor', 'house_fodor',
       'house_number_fodor', 'road_fodor', 'suburb_fodor', 'fodor_id',
       'match_status'],
      dtype='object')
```

As we can only match columns that were parsed in both address datasets, this means we lose two columns, `city_district_zagat` and `state_fodor`, that were parsed by the CRF segmentation model. Once we observe which address fields are common to both datasets, we create so-called comparison vectors from candidate address pairs of the Zagat and Fodor datasets. Each element of the comparison vector represents the string similarity between address fields contained in both databases. For example, `city_jaro` describes the string similarity between the columns `city_zagat` and `city_fodor`. Looking at the first two rows of our comparison vectors dataframe, a `city_jaro` value of 1.00 implies an exact match whereas a value of 0.4040 implies a number of modifications are required to match the two city names, and so these are less likely to correspond to a match.

[17]:
```
# create a function for building comparison vectors we can reuse later
def return_comparison_vectors(candidate_link_df):

   candidate_link_df['city_jaro'] = jarowinkler_similarity(candidate_link_df.
   ...city_zagat, candidate_link_df.city_fodor)
   candidate_link_df['house_jaro'] = jarowinkler_similarity(candidate_link_df.
   ...house_zagat, candidate_link_df.house_fodor)
   candidate_link_df['house_number_jaro'] = jarowinkler_similarity(candidate_link_df.
   ...house_number_zagat, candidate_link_df.house_number_fodor)
```

Table 3: Output of code 17

	city_jaro	house_jaro	house_number_jaro	road_jaro	suburb_jaro	match_status
0	1	1	1	1	0	1
1	0.40404	0.568301	0	0.629085	0	0
2	0	0.482143	0	0.674077	0	0
3	0.626263	0.502778	0.511111	0.629085	0	0
4	1	0.45463	0	0.831493	0	0

```
    candidate_link_df['road_jaro'] = jarowinkler_similarity(candidate_link_df.
    ...road_zagat, candidate_link_df.road_fodor)
    candidate_link_df['suburb_jaro'] = jarowinkler_similarity(candidate_link_df.
    ...suburb_zagat, candidate_link_df.suburb_fodor)

    # now we build a dataframe that contains the jaro-winkler similarity between the
    address components and the matching status
    comparison_vectors = candidate_link_df[['city_jaro', 'house_jaro',\
                                            'house_number_jaro','road_jaro',\
                                            'suburb_jaro', 'match_status']]

    # set NaN values to 0 so the comparison vectors can work with the applied classifiers
    comparison_vectors = comparison_vectors.fillna(0.)

    return comparison_vectors

comparison_vectors = return_comparison_vectors(candidate_link_df)

# lets preview this dataframe to build some intution as to how it looks
comparison_vectors.head().style.set_table_styles(styles)
```

[17]:
```
Output in Table 3
```

4.3 *Classification and evaluation of match performance*

Once we obtain comparison vectors for each candidate address pair, we frame our approach as a binary classification problem by resolving the vectors into matches and non-matches. As the Zagat and Fodors dataframe has labels that describe our address pairs as matched, we use supervised classification to train a statistical model, a **random forest**, to classify address pairs with an unknown match status into matches and non-matches. As a reminder, a random forest is generated using a multitude of decision trees during training which then outputs the mode of the match status decision for the individual trees.

In practice, we initialize a random forest object and split our `comparison_vectors` dataframe into features containing our Jaro-Winkler string similarity features, X, and a vector used to predict match status of the addresses, y.

[18]:
```
# create a random forest classifier that uses 100 trees and number of cores equal to
those available on machine
rf = RandomForestClassifier(n_estimators = 100,
                            # Due to small number of features (5) we do not limit
                            # depth of trees
                            max_depth = None,
                            # max number of features to evaluate split is
                            # sqrt(n_features)
                            max_features = 'auto',
                            n_jobs = os.cpu_count())

# define metrics we use to assess the model
scoring = ['precision', 'recall', 'f1']
folds = 10

# extract the jaro-winkler string similarity and match label
X = comparison_vectors.iloc[:, 0:5]
y = comparison_vectors['match_status']
```

To evaluate the performance of our built classification model, we use 10-fold cross-validation meaning the performance measures are averaged across the test sets used within the 10 folds. We use three metrics that are commonly used to evaluate machine learning models. Recall measures the proportion of address pairs that should have been classified, or recalled, as matched (Christen 2012). The precision (or, equivalently, the positive predictive value) calculates the proportion of the matched address pairs that are classified correctly as true matches (Christen 2012). Finally, the F1 score reflects the harmonic mean between precision and recall. Our cross-validation exercise is executed in the following cell.

[19]:
```
# 10-fold cross-validation procedure
scores = cross_validate(estimator = rf,
                X = X,
                y = y,
                cv = folds,
                scoring = scoring,
                return_train_score = False)
```

[20]:
```
print('Mean precision score is {} over {} folds.'.format( np.round(np.
...mean(scores['test_precision']), 4), folds))
print('Mean recall score is {} over {} folds.'.format( np.round(np.
...mean(scores['test_recall']), 4), folds))
print('Mean F1 score is {} over {} folds.'.format( np.round(np.
...mean(scores['test_f1']), 4), folds))
```

[20]:
```
Mean precision score is 0.9546 over 10 folds.
Mean recall score is 0.928 over 10 folds.
Mean F1 score is 0.9383 over 10 folds.
```

Overall, the high precision value implies that 95% of true positives are successfully disambiguated from false positives. Moreover, our recall value implies that 93% of all potential matches were successfully returned, with the remaining 7% of correct matches incorrectly labelled as false negatives. Given the high values in both of these metrics, the accompanying F1 score is equally high.

5 Creation of candidate address pairs by blocking on zipcode

While a Cartesian product could be useful in a linkage exercise where we have a very small number of matched addresses, in production environments more sophisticated techniques are generally required to create candidate address links. This is particularly the case when we have a large number of addresses. Thus, blocking is typically introduced to partition the set of all possible address comparisons to within mutually exclusive blocks. If we let b equal the number of blocks, we reduce the complexity of the comparison exercise to $O(\frac{n^2}{b})$, which is far more computationally tractable than the full index method used above.

When deciding which column to use as a blocking key we generally need pay attention to two main considerations. Firstly, we pay attention to attribute data quality. Typically when identifying a blocking key, we choose a key that has a **low number of missing values**. This is because choosing a key with many missing values forces a large number of addresses into a block where the key is an empty value, which may lead to many misclassified address matches. And secondly we pay attention to the **frequency distribution** of attribute values. We optimise towards a uniform distribution of values, as typically skewed distributions that result in some values occurring very frequently mean that these values will dominate the candidate pairs of address generated.

These considerations are addressed in the following two code blocks.

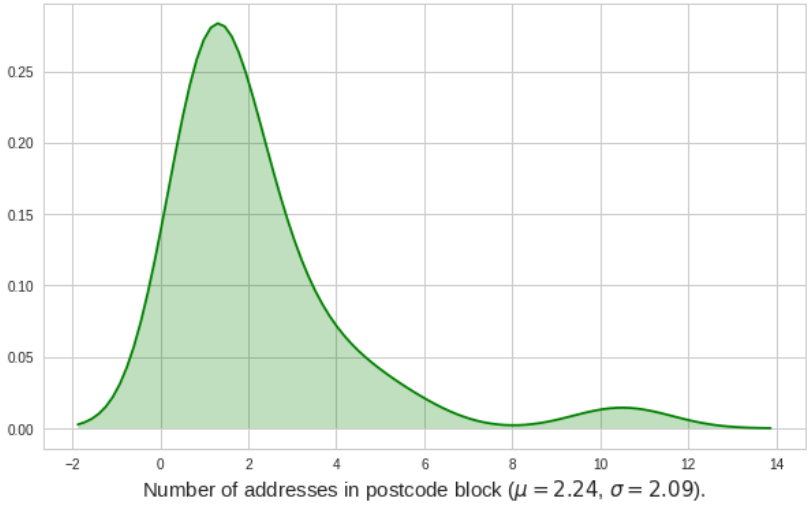


Figure 1: Figure generated by code 22

[21]:
```
print("Missing postcodes for Zagat addresses: {}. \n
Missing postcodes for Fodor addresses: {}.".format(matched_address.postcode_zagat.
...isnull().sum(), matched_address.postcode_fodor.isnull().sum()))
```

[21]:
```
Missing postcodes for Zagat addresses: 1.
Missing postcodes for Fodor addresses: 0.
```

[22]:
```
# check distribution of postcode blocks
pc_dist = matched_address.groupby('postcode_fodor').size().to_frame().
...rename(columns={0:'n_addresses'})

f, ax = plt.subplots(1, figsize=(10,6))
sns.kdeplot(pc_dist.n_addresses.values, color='g', shade=True, legend=False)
ax.set_xlabel('Number of addresses in postcode block ($\mu = {}$, $\sigma = {}$).'
              .format(np.mean(pc_dist.n_addresses), np.round(np.std(pc_dist.n_addresses),
              2)), size=15)
plt.show()
```

[22]:
```
Output in Figure 1
```

The postcode attribute looks like a sensible choice of blocking key because it contains just one missing value and there are very low numbers of candidate address comparisons within each block. As you can see in the output below, when we use a more sophisticated indexing technique we generate a far lower number of candidate address comparisons. In fact, we create only 1014 candidate address links despite adding synthetic non-matches (discussed below). Overall, our introduction of blocking substantially lowers the computational requirement of the linkage task.

5.1 Creation of synthetic non-matched addresses

To make this exercise more realistic, let's also create 112 synthetic non-matches so we have 224 addresses in total. This will also be important for training our machine learning technique to learn the representations of non-matched addresses in addition to matches. In this case we use the FEBRL data set generator script `generate.py` to create an artificially generated dataset (see http://users.cecs.anu.edu.au/ Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/node70.html). The script uses Python 2.7, so we read the output as JSON so the user does not have to rely on an external input. We do this in keeping with

a self-contained notebook but describe the steps required to reproduce the non-matches below.

The synthetic non-matches are essentially random permutations of the matched addresses. These are constructed on the basis of frequency tables for each address field that count the occurrence of particular values. For example, the first row of a frequency table for a house number would look like:

<house_number_attribute_value>,<frequency_of_occurence>.

[23]:
```
# first we need columns from the zagat and fodor databases to create random addresses
zagat_cols = ['city_zagat', 'house_number_zagat',\
              'house_zagat', 'suburb_zagat',\
              'road_zagat', 'postcode_zagat']
fodor_cols = ['city_fodor', 'house_number_fodor',\
              'house_fodor', 'suburb_fodor',\
              'road_fodor','postcode_fodor']

# create a directory for address component frequencies
if not os.path.exists('freqs'):
    os.makedirs('freqs')

# create distributions of address components for both datasets that will be used to
# create fake addresses
for cols in [zagat_cols, fodor_cols]:
    for col in cols:
        freq = matched_address[col].value_counts().reset_index()
        freq.to_csv('freqs/{}_freq.csv'.format(col), index=False, header=False)
```

The script `generate.py` takes six parameters that are used to create non-matched addresses. The first argument demarcates the number of original records to be generated; the second specifies the number of duplicate records from the original to be generated; and the third, fourth and fifth arguments define the maximal number of duplicate records that will be created based on one original record, the maximum number of modifications introduced to the address field, and the maximum number of modifications introduced to the address, respectively. The final parameter is used to enter which probability distribution will be to create duplicate records – i.e. uniform, poisson, or zipf. In our case we are only interested in building synthetic non-matches (and not duplicates), so we set the number of original records to be built as 112, the number of duplicates generated as 0, and leave the number of modifications introduced by the recommended default settings.

In addition, for each address field, users are asked to define a dictionary inside `generate.py` that outlines the probability for particular modifications. This includes setting the probability of modifications such as misspellings, insertions, deletions, substitutions and transpositions of word and characters. An example dictionary for the house number address field is given below where we set the file path to the word frequency CSV generated above:

[24]:
```
house_number_dict = {'name':'house_number',
                     'type':'freq',
                'char_range':'digit',
               # 'freq_file':'freqs/house_number_fodor_freq.csv',
                'freq_file':'freqs/house_number_zagat_freq.csv',
               'select_prob':0.20,
                  'ins_prob':0.10,
                  'del_prob':0.16,
                  'sub_prob':0.54,
                'trans_prob':0.00,
             'val_swap_prob':0.00,
             'wrd_swap_prob':0.00,
              'spc_ins_prob':0.00,
              'spc_del_prob':0.00,
                 'miss_prob':0.00,
              'new_val_prob':0.20}
```

Damerau (1964) finds the proportions of typographical errors are typically spread as substitutions (59%), deletions (16%), transpositions (2%), insertions (10%) and multiple errors (13%). For this reason we broadly align our dictionary probabilities with these findings. After defining sensible probabilities for modifications, we execute the following scripts on a terminal which will create a file, `zagat_synthetic_addresses.csv` and `fodor_synthetic_addresses.csv` consisting of synthetic addresses from the Zagat and Fodor datasets, respectively.

For simplicity we generate our non-matches using all the data at once. However, in a real-world application, we might wish to create non-matches within each zipcode block one at a time. This would create more realistic synthetic non-matches. This is because non-matched addresses would be constructed from the frequency tables of each zipcode block, meaning each non-match would share more commonality to actual matched addresses. In practice, this would improve the predictive power of our classification model to disambiguate between candidate address pairs that have very subtle differences yet are still matched or non-matched.

[25]:
```
# ! python2 generate.py zagat_synthetic_addresses.csv 112 0 4 2 2 poisson
```

[25]:
```
Create 112 original and 0 duplicate records
  Distribution of number of duplicates (maximal 4 duplicates):
  [(1, 0.0), (2, 0.375), (3, 0.75), (4, 0.9375)]

Step 1: Load and process frequency tables and misspellings dictionaries

Step 2: Create original records


Step 2: Create duplicate records


Step 3: Write output file
End.
```

[26]:
```
# ! python2 generate.py fodor_synthetic_addresses.csv 112 0 4 2 2 poisson
```

[26]:
```
Create 112 original and 0 duplicate records
  Distribution of number of duplicates (maximal 4 duplicates):
  [(1, 0.0), (2, 0.375), (3, 0.75), (4, 0.9375)]

Step 1: Load and process frequency tables and misspellings dictionaries

Step 2: Create original records


Step 2: Create duplicate records


Step 3: Write output file
End.
```

We then read these synthetic non-matches into a dataframe.

[27]:
```
# read parsed synthetic addresses
synthetic_zagat_address = pd.read_csv('zagat_synthetic_addresses.csv').
...add_suffix('_zagat').drop(columns=['rec_id_zagat'])
synthetic_fodor_address = pd.read_csv('fodor_synthetic_addresses.csv').
...add_suffix('_fodor').drop(columns=['rec_id_fodor'])
```

```
# set uids for synthetic addresses
synthetic_zagat_address['zagat_id'] = [str(uuid.uuid4()) for i in
...synthetic_zagat_address.iterrows()]
synthetic_fodor_address['fodor_id'] = [str(uuid.uuid4()) for i in
...synthetic_fodor_address.iterrows()]

# join synthetic zagat and fodor addresses vertically
synthetic_non_matches = synthetic_zagat_address.join(synthetic_fodor_address)

# remove whitespace from column names and attributes
synthetic_non_matches = synthetic_non_matches.rename(columns = lambda x : x.strip())
synthetic_non_matches = synthetic_non_matches.applymap(lambda x : x.strip() if
...type(x) == str else x)
```

Now we have generated synthetic non-matches, we need to join these back to our dataframe of matched addresses. As the above steps require external scripts we provide the JSON required to reconstruct the synthetic dataframe in the dedicated Github repository. This can be read by executing the cell below which uses the `pd.read_json` function.

[28]:
```
! wget https://raw.githubusercontent.com/SamComber/address_matching_workflow/master/
...synthetic_addresses.json
```

[28]:
```
--2019-12-21 09:11:11--  https://raw.githubusercontent.com/SamComber/address_matching_
workflow/master/synthetic_addresses.json
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 199.232.56.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|199.232.56.133|:443
... connected.
HTTP request sent, awaiting response... 200 OK
Length: 29098 (28K) [text/plain]
Saving to: ‘synthetic_addresses.json’

synthetic_addresses 100%[===================>]  28.42K  --.-KB/s    in 0.02s

2019-12-21 09:11:11 (1.16 MB/s) - ‘synthetic_addresses.json’ saved [29098/29098]
```

[29]:
```
f = 'synthetic_addresses.json'

synthetic_non_matches = pd.read_json(f)
```

In the cell below we join our matched addresses with our synthetic non-matches, creating a dataframe of 224 address pairs.

[30]:
```
# align columns of matched_address dataframe for horizontal join
matched_address = matched_address[['house_zagat', 'house_number_zagat', 'road_zagat',
... 'suburb_zagat', 'city_zagat', 'postcode_zagat','zagat_id', 'house_fodor',
... 'house_number_fodor', 'road_fodor', 'suburb_fodor', 'city_fodor', 'postcode_fodor',
... 'fodor_id']]

# horizontal join between matched addresses and synthetic non-matches
matches_with_non_matches = pd.concat([matched_address, synthetic_non_matches],
... ignore_index=True)

print('{} address pairs created consisting of {} matches and {} synthetic non-matches.'.
...format(matches_with_non_matches.shape[0], matched_address.shape[0],
...       synthetic_non_matches.shape[0]))
```

[30]:
```
224 address pairs created consisting of 112 matches and 112 synthetic non-matches.
```

5.2 *Blocking on postcode attribute*

With our matches and synthetic non-matches assembled into a dataframe with 224 address pairs, we can proceed to block on postcode values to create mutually exclusive address partitions. Thus, for every unique postcode value, a dataframe (or block) will be created in which candidate address pairs will be matched and non-matched based on attributes of their comparison vectors.

The following code block creates a `MultiIndex` that links together the IDs of addresses that are within the same zipcode block.

[31]:
```
indexer = rl.Index()

# block on postcode attribute
indexer.block(left_on='postcode_zagat', right_on='postcode_fodor')
candidate_links = indexer.index(matches_with_non_matches, matches_with_non_matches)

# this creates a two-level multiindex, so we name addresses from the zagat and fodor
databases, respectively.
candidate_links.names = ['zagat', 'fodor']

print('{} candidate links created using the postcode attribute as a blocking key.'.
...format(len(candidate_links)))
```

[31]:
```
1014 candidate links created using the postcode attribute as a blocking key.
```

We follow the same work flow as before and create comparison vectors for every 1014 candidate address links.

[32]:
```
candidate_link_df = return_candidate_links_with_match_status(candidate_links)

comparison_vectors = return_comparison_vectors(candidate_link_df)
```

Following this, we train our random forest on the comparison vectors and match status labels. We use a 75/25 split for our train and test data.

[33]:
```
X = comparison_vectors.iloc[:, 0:5]
y = comparison_vectors.match_status

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)

# create a random forest classifier that uses 100 trees and number of cores equal to
those available on machine
rf = RandomForestClassifier(n_estimators = 100,
                            # Due to small number of features (5) we do not limit
                            # depth of trees
                            max_depth = None,
                            # max number of features to evaluate split is
                            # sqrt(n_features)
                            max_features = 'auto',
                            n_jobs = os.cpu_count())

# predict match status of unseen address pairs
y_pred = rf.fit(X_train, y_train).predict(X_test)
```

5.3 *Classification and evaluation of match performance*

Having fit our random forest on the training data we can now assess the model under the number of metrics we introduced earlier. We can also produce a confusion matrix which shows true negatives in the top-left quadrant, false positives in the top-right, false negatives in the bottom-left and true positives in the bottom-right. At first glance, the

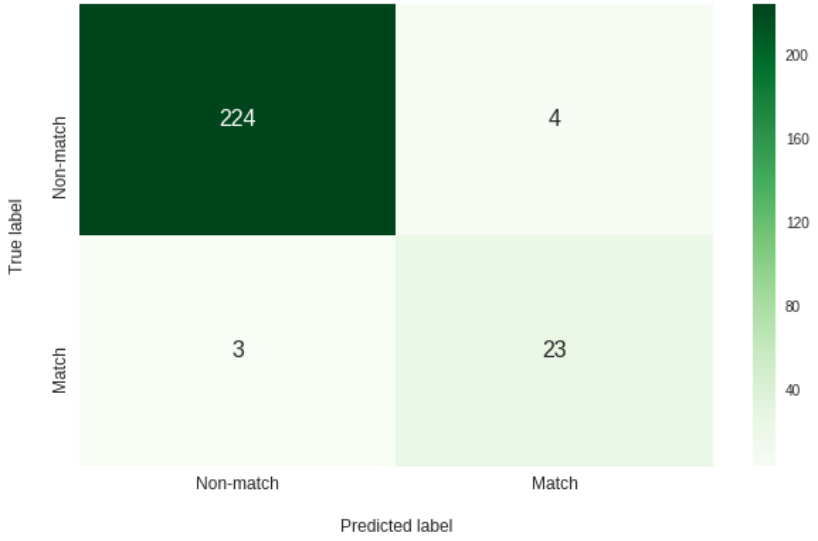


Figure 2: Figure generated by code 34

findings from the evaluation metrics below may seem counter-intuitive, especially as the results of the classification exercise using the full index performed better. However, it is pertinent to remind ourselves that we trained our classification model on **matched address only**, which reflected an idealised but unrealistic scenario. In the results below we introduced synthetic non-matches which reflected a scenario that a user is more likely to encounter in a real-world address matching exercise.

In the following code block we generate evaluation metrics and a confusion matrix for evaluating match performance.

[34]:
```
print('Precision score: {}.'.format(np.round(precision_score(y_test, y_pred), 4)))
print('Recall score: {}.'.format(np.round(recall_score(y_test, y_pred), 4)))
print('F1 score: {}.'.format(np.round(f1_score(y_test, y_pred), 4)))

f,ax = plt.subplots(1, figsize=(10,6))
f.set_tight_layout(False)

fontsize=12
sns.heatmap(confusion_matrix(y_test, y_pred),
            ax=ax,
            annot=True,
            annot_kws={'fontsize': 16},
            cmap='Greens',
            fmt='g')
ax.set_yticklabels(['Match', 'Non-match'], fontsize=fontsize)
ax.set_xticklabels(['Non-match', 'Match'], fontsize=fontsize);
ax.set_ylabel('True label', fontsize=fontsize)
ax.set_xlabel('Predicted label', fontsize=fontsize)
ax.xaxis.labelpad = 18
ax.yaxis.labelpad = 18
plt.show();
```

[34]:
```
Precision score: 0.8519.
Recall score: 0.8846.
F1 score: 0.8679.

Output in Figure 2
```

Overall, our precision value implied 85% of true positives were correctly separated from false positives, and our recall value indicated that 88% of all true address matches were successfully retrieved, with the remaining 12% incorrectly classified as non-matches.

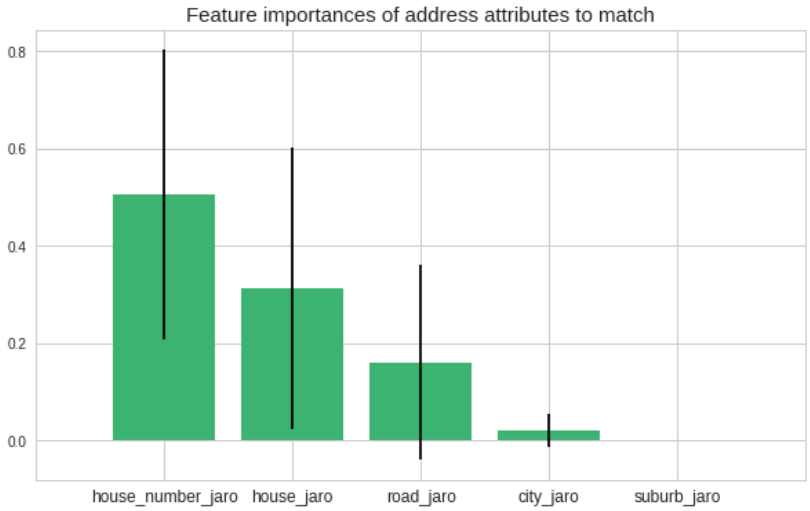


Figure 3: Output generated by code 35

With our model now fitted and tested, we could extend its use to predict the match status of unseen address pairs. As an example application, if we had a small sample of matched addresses that belonged to a larger set of unmatched addresses, we could use our trained predictive model to match the remaining addresses in the dataset. This would work so long as the textual representations of addresses used in the prediction stage follow a similar structure to those addresses used to train the classification model.

Before we conclude, a benefit of using ensemble methods such as random forest classifiers is that we can return an indication of how useful and valuable each feature was in the construction of each decision tree. In a practical application, extracting a measure of feature importance might be a useful step in pruning redundant features from the comparison vectors. This might be a useful step in lowering computation times as we decrease the number of address field comparisons required to evaluate candidate address pairs.

Thus, in the following code block we rank feature importance of particular address fields to the match classification.

[35]:
```
# extract feature importances from random forest classifier
feature_importance_to_match = rf.feature_importances_

# calculate standard deviation of feature importances across trees
std = np.std([tree.feature_importances_ for tree in rf.estimators_], axis=0)
indices = np.argsort(feature_importance_to_match)[::-1]

# plot importances alongside feature labels
plt.figure(figsize=(10,6))
plt.title("Feature importances of address attributes to match", size=15)
plt.bar(range(X_train.shape[1]), feature_importance_to_match[indices],
       color="#3CB371", yerr=std[indices], align="center")
feature_labs = X_train.columns[np.argsort(feature_importance_to_match)[::-1]].values
plt.xticks(range(X_train.shape[1]), feature_labs, size=12)
plt.xlim([-1, X_train.shape[1]])
plt.show()
```

[35]:
```
Output in Figure 3
```

In our case, and as one might expect, the restaurant's house number, `house_number_jaro` is the most important feature used for resolving candidate pairs of addresses into a match while the suburb, `suburb_jaro`, is the least important feature and so could possibly be removed as an address field from the comparison step.

6 Conclusion

Address matching is a data enrichment process that is increasingly required in wide-ranging, real-world applications. For example, matching between census, commercial or lifestyle records has the potential benefit of improving data quality, enabling spatial data visualisation and joining data that would otherwise remain isolated in data silos. In absence of unique identifiers for directly linking data, practitioners have typically relied on statistical linkage methods for matching addresses. Linking address datasets in this way has the potential to unlock attributes that one would be unable to access in circumstances where no primary keys exist to join the two datasets. Thus, in this notebook, we documented the steps required to execute the work flow for an address matching exercise that utilised new and recent innovations in machine learning. While the dataset we used was low volume, the intention of the notebook was to demonstrate an approach that is reproducible within a self-contained environment, and which might be adapted by the interested user to larger data challenges. Training a predictive model to link restaurant addresses may seem a trivial problem to solve, but these addresses could easily be replaced by more meaningful address records in areas such as public health and socio-economic mobility studies. Therefore, the core contribution of this notebook sought to equip the regional scientist with skills necessary to extend the address matching work flow to their own (and far more interesting) use cases.

References

Baldovin T, Zangrando D, Casale P, Ferrarese F, Bertoncello C, Buja A, Marcolongo A, Baldo V (2015) Geocoding health data with geographic information systems: A pilot study in northeast Italy for developing a standardized data-acquiring format. *Journal of Preventive Medicine & Hygiene* 56: 88–94

Cayo R, Talbot TO (2003) Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2: 1–10

Christen P (2012) *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer, New York, NY

Comber S, Arribas-Bel D (2019) Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. *Transactions in GIS* 23: 334–348

Damerau F (1964) A technique for computer detection and correction of spelling errors. *Commununications of the ACM* 7: 171–176. CrossRef.

Diesner J, Carley M (2008) Conditional random fields for entity extraction and ontological text coding. *Computational and Mathematical Organization Theory* 14: 248–262

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In: Brodley CE, Danyluk AP (eds), *Proceedings of the 18th International Conference on Machine Learning.* Morgan Kaufmann, San Francisco, CA, 282–289

Reynolds P, Behren JV, Gunier R, Goldberg D, Hertz A, Smith D (2003) Childhood cancer incidence rates and hazardous air pollutants in California: An exploratory analysis. *Environmental Health Perspectives* 111: 663–668

Ruggles S, Fitch C, Roberts E (2018) Historical census record linkage. *Annual Review of Sociology* 44[1]: 19–37

Yancey W (2005) Evaluating string comparator performance for record linkage. Research report series, statistics #2005-05, Bureau of the Census, Washington, DC

REGION
The Journal of ERSA
Powered by WU

Volume 6, Number 3, 2019, 39–51
DOI: 10.18335/region.v6i3.278

journal homepage: region.ersa.org

Urban Street Network Analysis in a Computational Notebook*

Geoff Boeing[1]

[1] University of Southern California, Los Angeles, USA

Received: 21 September 2019/Accepted: 20 December 2019

Abstract. Computational notebooks offer researchers, practitioners, students, and educators the ability to interactively conduct analytics and disseminate reproducible workflows that weave together code, visuals, and narratives. This article explores the potential of computational notebooks in urban analytics and planning, demonstrating their utility through a case study of OSMnx and its tutorials repository. OSMnx is a Python package for working with OpenStreetMap data and modeling, analyzing, and visualizing street networks anywhere in the world. Its official demos and tutorials are distributed as open-source Jupyter notebooks on GitHub. This article showcases this resource by documenting the repository and demonstrating OSMnx interactively through a synoptic tutorial adapted from the repository. It illustrates how to download urban data and model street networks for various study sites, compute network indicators, visualize street centrality, calculate routes, and work with other spatial data such as building footprints and points of interest. Computational notebooks help introduce methods to new users and help researchers reach broader audiences interested in learning from, adapting, and remixing their work. Due to their utility and versatility, the ongoing adoption of computational notebooks in urban planning, analytics, and related geocomputation disciplines should continue into the future.

Key words: Computational Notebook, Jupyter, OpenStreetMap, OSMnx, Python, Street Network, Urban Planning

1 Introduction

A traditional academic and professional divide has long existed between code creators and code users. The former would develop software tools and workflows for professional or research applications, which the latter would then use to conduct analyses or answer scientific questions. Today, however, these boundary lines increasingly blur as computation percolates throughout both the natural and social sciences. As quantitatively-oriented academics gradually shift away from monolithic, closed-source data analysis software systems like SPSS and ArcGIS, they increasingly embrace coding languages like R and Python to script and document their research workflows (Padgham et al. 2019). Developing shareable, reproducible, and recomputable scripts in R or Python to acquire, transform, describe, visualize, and model data, these researchers act as both code creators and code users.

*This paper is available as computational notebook on the REGION webpage.

An important trend in this methodological trajectory has been the widespread adoption of the computational notebook. A computational notebook is a computer file that replaces the traditional lab notebook and intersperses plain-language narrative, hyperlinks, and images with snippets of code in the paradigm of literate programming (Knuth 1992). These notebooks are easily distributed and integrate well with version control systems like Git because they are simply structured text files. They have pedagogical value in introducing students to computational thinking and coding techniques while thoroughly explaining each new programming language facet as it is introduced. They also offer research value in documenting data, questions, hypotheses, procedures, experiments, and results in detail alongside each's attendant computations (Pérez, Granger 2007, Kluyver et al. 2016).

Computational notebooks thus open up the world of analytics to a wider audience than was possible in the past. This particularly impacts disciplines that encompass diverse methodologies and skillsets. For example, urban planning, like many academic domains related or adjacent to regional science, comprises a broad set of scholars, students, and working professionals with a wide range of computational aptitude. Some urban planners focus on policymaking within the political constraints of city hall. Others employ qualitative methods to work in and with vulnerable communities. Others develop simulation models to forecast urbanization patterns and infrastructure needs. Others intermingle these, and many more, different approaches to understanding and shaping the city. Yet all urban planners benefit from basic quantitative literacy and an ability to reason critically with data. This scholarly and professional imperative aligns with the growing importance of computational thinking in the urban context and parallel trends in geocomputation (Harris et al. 2017), geographic data science (Kang et al. 2019, Poorthuis, Zook 2019, Singleton, Arribas-Bel 2019), and the open-source/open-science movements (Rey 2019).

Urban planning and its related disciplines benefit accordingly from the growing adoption of computational notebooks in pedagogy, research, and practice. Computation is increasingly central to the field and its practitioners benefit from open and reproducible approaches to analyzing urban data and predicting city futures (Kedron et al. 2019, Kontokosta 2018, Batty 2019). In the Python universe, for example, numerous new tools now exist to support urban analytics and planning processes, including data wrangling/analysis (pandas), visualization (matplotlib), geospatial wrangling/analysis (geopandas), spatial data science and econometrics (pySAL), mapping (cartopy), web mapping (folium), network analysis (NetworkX), land use modeling/simulation (UrbanSim), activity-based travel modeling (ActivitySim), and computational notebooks themselves (Jupyter).

Another Python tool useful for urban planning research and practice – and the primary focus of this article – is OSMnx, a package for street network analysis (Boeing 2017). OSMnx allows users to download spatial data (including street networks, other networked infrastructure, building footprints, and points of interest) from OpenStreetMap then model, analyze, and visualize them. To introduce new users to its functionality and capabilities, OSMnx's official demos and tutorials are developed and maintained in Jupyter notebook format. This repository in turn offers a compelling case study of the potential of computational notebooks to document and disseminate geospatial software tools.

This article introduces OSMnx as a computational tool for urban street network analysis by way of these computational notebooks. It describes their repository and highlights examples from them, inline here, to illustrate the use and value of computational notebooks. To do so, it demonstrates how to interactively execute the code in this article itself by using Docker to run a containerized computational environment including Jupyter Lab as an interactive web-based interface. The article is organized as follows. First, it presents the repository containing OSMnx's demo and tutorial notebooks. Then it describes how to run OSMnx's computational environment via Docker. Next it demonstrates the use of OSMnx interactively in the article itself through a synoptic tutorial adapted from this repository. Finally, it concludes by discussing the prospects of notebooks for facilitating the adoption of computational workflows in urban analytics and planning.

2 The OSMnx Examples Repository

OSMnx's official demos, tutorials, and examples are in Jupyter notebook format in a GitHub repository. The repository's root contains a license file, a readme file, an environment definition file, repository contributing guidelines, and a notebooks folder. Within that folder, the repository contains 19 thematically organized Jupyter notebook files that collectively provide a short self-directed tutorial-style course in using OSMnx. The following notebooks are included there:

1. An introductory survey of features
2. A more comprehensive overview of OSMnx's basic functionality
3. Using OSMnx to produce shapefiles
4. Modeling and visualizing street networks in different places at different scales
5. Using OSMnx's network topology cleaning and simplification features
6. Saving and loading data to/from disk with OSMnx
7. Conducting street network analyses with OSMnx and its NetworkX dependency
8. Visualizing street networks and study sites
9. Working with dual graphs of street networks
10. Producing figure-ground diagrams for urban form analysis
11. Working with building footprints
12. Interactive web mapping of street networks and routes
13. Attaching elevations to the network and calculating street grades
14. Working with isolines and isochrones
15. Cleaning complex street intersections
16. Calculating street bearings
17. Working with other types of spatial infrastructure
18. Visualizing street network orientation with polar histograms
19. Interfacing between OSMnx and igraph for fast algorithm implementations in the C language

This resource is useful for introducing users to the OSMnx software package, demonstrating how to download, model, analyze, and visualize street networks in Python, and illustrating several basic and intermediate spatial network analyses. To run the code examples in this resource repository, one must have access to a Python installation with the code dependencies installed, including Jupyter itself for running the notebook files. Two primary options exist for installing this computational environment. The first is installing Python locally, then configuring it and installing all the necessary packages and dependencies. This can be time-consuming and requires some prior experience beyond the scope of this article. The second, and easier, option is to simply run everything in a pre-built Docker container. This latter option is detailed in the following section.

3 The Computational Environment

The OSMnx project's reference Docker image contains a stable, consistent computational environment for running OSMnx on any computer. Docker is a virtualization tool that allows complex software stacks to be delivered as self-contained packages called images, allowing users to run software without having to compile or install a complex chain of dependencies. Instead, users install Docker on their computer then tell it to run a certain image as an instance called a container.

This article can be read in its static form (i.e., HTML or PDF) or it can be executed interactively (i.e., via its .ipynb Jupyter notebook file). For interactive execution, install Docker and run the official OSMnx container as follows. First, download and install Docker Desktop. Once it is installed and running on your computer, open Docker's settings/preferences and ensure that your local drives are shared with Docker so the container has access to the notebook file. Then run the OSMnx Docker container (which contains a Python installation and all the packages needed to run OSMnx, including Jupyter Lab) by following the platform-specific instructions below.

If you are on *Windows* open a command prompt, change directory to the location of this notebook file then run:

```
docker run --rm -it -p 8888:8888 -v "\%cd\%":/home/jovyan/work gboeing/osmnx:v10
```

If you are on *Mac/Linux* open a terminal window, change directory to the location of this notebook file then run:

```
docker run --rm -it -p 8888:8888 -v "$PWD":/home/jovyan/work gboeing/osmnx:v10
```

Once the container is running per these instructions, open your computer's web browser and visit http://localhost:8888 to access Jupyter Lab and open this article's notebook file.

4 Street Network Analysis with OSMnx

Here we showcase the resource repository inline to demonstrate potential applications. In particular, we highlight specific material from its notebooks (enumerated above), adapting their code into this interactive article to introduce OSMnx and illustrate some of the capabilities of a computational notebook.

First we import the necessary Python modules:

[1]:
```
import matplotlib.cm as cm
import matplotlib.colors as colors
import networkx as nx
from IPython.display import Image
from pprint import pprint
```

matplotlib is a package for data visualization and plotting. NetworkX is a package for generic network analysis. IPython provides interactive computing and underpins our Python-language Jupyter environment (Pérez, Granger 2007). pprint allows us to "pretty print" Python data structures to make them easier to read inline.

Next we import OSMnx itself, configure it, and display its version number:

[2]:
```
import osmnx as ox
ox.config(log_console=True, use_cache=True)
ox.__version__
```

[2]:
```
'0.10'
```

The configuration step tells OSMnx to log its actions to the terminal window and to use a cache. This cache saves a local copy of any data downloaded by OSMnx to prevent re-downloading the same data each time the code is run.

Next we use OSMnx to download the street network of Piedmont, California, construct a graph model of it (via NetworkX), then plot the network with the `plot_graph` function (which uses matplotlib under the hood):

[3]:
```
# create a graph of Piedmont's drivable street network then plot it
G = ox.graph_from_place('Piedmont, California, USA', network_type='drive')
fig, ax = ox.plot_graph(G)
```

[3]:
```
For the output see Figure 1
```

In the resulting Figure 1, the network's intersections and dead-ends (i.e., graph nodes) appear as light blue circles and its street segments (i.e., graph edges) appear as gray lines. This is the street network within the municipal boundaries of the city of Piedmont, California. We select this study site for pedagogical purposes as it is a relatively small, self-contained municipality and lends itself to convenient visualization and indicator calculation here. Note that we specified `network_type='drive'` so this is specifically the drivable network in the city. OSMnx can also automatically download and model walkable and bikeable street networks by changing this argument.

4.1 Calculating Network Indicators

Now that we have a model of the network, we can calculate some statistics and indicators. First, what area does our network cover in square meters? To calculate this, we project the graph, convert its projected nodes to a geopandas GeoDataFrame, then calculate the area of the convex hull of this set of node points in the Euclidean plane:

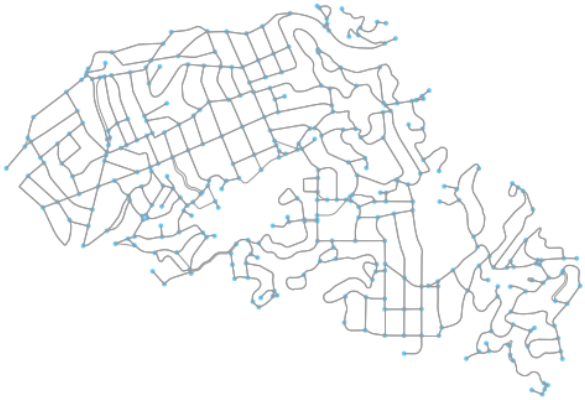


Figure 1: Output from codebox 3

[4]:
```
# project graph then calculate its nodes' convex hull area
G_proj = ox.project_graph(G)
nodes_proj = ox.graph_to_gdfs(G_proj, edges=False)
graph_area_m = nodes_proj.unary_union.convex_hull.area
graph_area_m
```

[4]:
```
4224782.349449131
```

Thus, this network covers approximately 4.2 square kilometers. When projecting graphs, OSMnx by default uses the Universal Transverse Mercator (UTM) coordinate system and automatically determines the UTM zone for projection based on the network's centroid. Other coordinate reference systems can be defined by the user to customize this projection behavior.

Next, we compute and inspect some basic stats about the network:

[5]:
```
# calculate and print basic network stats
stats = ox.basic_stats(G_proj, area=graph_area_m, clean_intersects=True,
        circuity_dist='euclidean')
pprint(stats)
```

[5]:
```
{'circuity_avg': 1.11354525174028,
 'clean_intersection_count': 271,
 'clean_intersection_density_km': 64.1453162753664,
 'edge_density_km': 26951.828421373437,
 'edge_length_avg': 121.39190724946685,
 'edge_length_total': 113865.60899999991,
 'intersection_count': 312,
 'intersection_density_km': 73.84995822108604,
 'k_avg': 5.421965317919075,
 'm': 938,
 'n': 346,
 'node_density_km': 81.89771007851208,
 'self_loop_proportion': 0.006396588486140725,
 'street_density_km': 14061.652905680734,
 'street_length_avg': 121.23963877551029,
 'street_length_total': 59407.42300000004,
 'street_segments_count': 490,
 'streets_per_node_avg': 2.953757225433526,
 'streets_per_node_counts': {0: 0, 1: 34, 2: 0, 3: 263, 4: 47, 5: 1, 6: 1},
 'streets_per_node_proportion': {0: 0.0,
                                 1: 0.09826589595375723,
                                 2: 0.0,
                                 3: 0.7601156069364162,
                                 4: 0.13583815028901733,
                                 5: 0.002890173410404624,
                                 6: 0.002890173410404624}}
```

For example, we can see that this network has 346 nodes (n) and 938 edges (m). The

streets in this network are 11% more circuitous (*circuity_avg*) than straight-line would be. The average street segment length is 121 meters (*street_length_avg*). We can inspect more stats, primarily topological in nature, with the `extended_stats` function. As the results of many of these indicators are verbose (i.e., calculated at the node-level), we print only the indicators' names here:

[6]:
```
# calculate and print extended network stats
more_stats = ox.extended_stats(G, ecc=True, bc=True, cc=True)
for key in sorted(more_stats.keys()):
    print(key)
```

[6]:
```
avg_neighbor_degree
avg_neighbor_degree_avg
avg_weighted_neighbor_degree
avg_weighted_neighbor_degree_avg
betweenness_centrality
betweenness_centrality_avg
center
closeness_centrality
closeness_centrality_avg
clustering_coefficient
clustering_coefficient_avg
clustering_coefficient_weighted
clustering_coefficient_weighted_avg
degree_centrality
degree_centrality_avg
diameter
eccentricity
pagerank
pagerank_max
pagerank_max_node
pagerank_min
pagerank_min_node
periphery
radius
```

The average neighborhood degree indicators refer to the mean degree of nodes in the neighborhood of each node. The centrality indicators (betweenness, closeness, degree, and PageRank) identify how "central" or important each node is to the network in terms of its topological structure. The clustering coefficient indicators represent the extent to which a node's neighborhood forms a complete graph. The extended stats also include the network's eccentricity (the maximum distance from each node to all other nodes), diameter (maximum eccentricity in the network), radius (minimum eccentricity in the network), center (set of all nodes whose eccentricity equals the radius), and periphery (set of all nodes whose eccentricity equals the diameter). Additional information about the various indicators is available online in OSMnx's documentation.

Now that we have modeled the street network and computed various indictors of its geometry and topology, we can finally save our graph to disk as an ESRI shapefile or a GraphML file (an open-source format for graph serialization), allowing easy re-use in other GIS or network analysis software:

[7]:
```
# save the network model to disk as a shapefile and graphml
ox.save_graph_shapefile(G, filename='mynetwork_shapefile')
ox.save_graphml(G, filename='mynetwork.graphml')
```

4.2 Visualizing Street Centrality

OSMnx is built on top of NetworkX, a powerful network analysis package developed at Los Alamos National Laboratory (Hagberg et al. 2008). We can use it to calculate and visualize the closeness centrality of different streets in the network. Closeness centrality measures how central a node or edge is in a network and is defined as the reciprocal of the sum of the distance-weighted shortest paths between the node/edge and every other node/edge in the network.

First, we convert our graph to its line graph (sometimes called the *dual graph*; see Porta et al. 2006) which inverts its topological definitions such that streets become nodes

Figure 2: Output from codebox 9

and intersections become edges. Then we calculate the closeness centrality of each node (i.e., street in the line graph):

[8]:
```
# calculate node closeness centrality of the line graph
edge_centrality = nx.closeness_centrality(nx.line_graph(G))
```

Now that we have calculated the centrality of each street in the network, we visualize it with matplotlib via OSMnx's `plot_graph` function, using the inferno color map to represent the most-central streets in bright yellow and the least-central streets in dark purple (see Figure 2):

[9]:
```
# make a list of graph edge centrality values
ev = [edge_centrality[edge  (0,)] for edge in G.edges()]

# create a color scale converted to list of colors for graph edges
norm = colors.Normalize(vmin=min(ev)*0.8, vmax=max(ev))
cmap = cm.ScalarMappable(norm=norm, cmap=cm.inferno)
ec = [cmap.to_rgba(cl) for cl in ev]

# color the edges in the original graph by closeness centrality in line graph
fig, ax = ox.plot_graph(G, bgcolor='black', axis_off=True, node_size=0,
                        edge_color=ec, edge_linewidth=2, edge_alpha=1)
```

[9]:
```
For the output see Figure 2
```

4.3 Network Routing

OSMnx allows researchers and practitioners to calculate routes and simulate trips along the network using various shortest-path algorithms, such as that by Dijkstra (1959). We demonstrate this here. First we use OSMnx to find the network nodes nearest to two latitude-longitude points:

[10]:
```
# find the network nodes nearest to two points
orig_node = ox.get_nearest_node(G, (37.825956, -122.242278))
dest_node = ox.get_nearest_node(G, (37.817180, -122.218078))
```

Next we compute the shortest path between these origin and destination nodes using Dijkstra's algorithm weighted by length (i.e., geometric distance along the street network). Then we use OSMnx to plot this route along the network:

[11]:
```
# calculate the shortest path between these nodes then plot it
route = nx.shortest_path(G, orig_node, dest_node, weight='length',
        method='dijkstra')
fig, ax = ox.plot_graph_route(G, route, node_size=0)
```

[11]:
```
For the output see Figure 3
```

Finally, we can calculate some statistics of our route, including its total length, in meters:

Figure 3: Output from codebox 11

[12]:
```
# what is the network distance of this route?
net_dist = nx.shortest_path_length(G, orig_node, dest_node, weight='length',
           method='dijkstra')
net_dist
```

[12]:
```
3284.0989999999997
```

Thus, this trip would travel approximately 3.3 kilometers along the network. We can also calculate the straight-line distance between these two network nodes as-the-crow-flies, using OSMnx's vectorized great-circle calculator:

[13]:
```
# what is the straight-line distance from origin to destination?
sl_dist = ox.great_circle_vec(G.node[orig_node]['y'], G.node[orig_node]['x'],
                              G.node[dest_node]['y'], G.node[dest_node]['x'])
sl_dist
```

[13]:
```
2340.8766018171827
```

Comparing these two distance values, we can compute an indicator of trip circuity: that is, how much greater the network-constrained distance is between two nodes compared to the straight-line distance between them. In this case, we can see that the network distance is approximately 40% longer than the straight-line distance:

[14]:
```
# how much longer is the network distance than the straight-line?
net_dist / sl_dist
```

[14]:
```
1.4029355487814306
```

4.4 Downloading/Modeling Networks in Other Ways

So far, we have modeled and analyzed the street network of Piedmont, California. However, we are not constrained to study sites in the United States. OpenStreetMap is a global mapping project and OSMnx can model networks anywhere in the world, such as Modena, Italy:

[15]:
```
# create a graph of Modena's drivable street network then plot it
G = ox.graph_from_place('Modena, Italy', retain_all=True)
fig, ax = ox.plot_graph(G, fig_height=8, node_size=0, edge_linewidth=0.5)
```

[15]:
```
For the output see Figure 4
```

We have seen how to download street network data and turn it into a graph-based model using OSMnx's `graph_from_place` function. This function geocodes the place name using OpenStreetMap's Nominatim web service, identifies its bounding polygon, then downloads all the network data within this polygon from OpenStreetMap's Overpass API. This workflow easily handles well-defined place names. However, OSMnx offers additional functionality to download and model networks for other study sites as well.

For example, if OpenStreetMap does not have a bounding polygon for a specific study site, we can acquire its street network anyway by passing a polygon directly into the

Figure 4: Output from codebox 15

`graph_from_polygon` function. Or we can pass in latitude-longitude coordinates and a distance into the `graph_from_point` function as demonstrated here, where we visualize the network within a bounding box around the University of California, Berkeley's Wurster Hall:

[16]:
```
# create a graph around UC Berkeley then plot it
wurster_hall = (37.870605, -122.254830)
one_mile = 1609 #one mile in meters
G = ox.graph_from_point(wurster_hall, distance=one_mile, network_type='drive')
fig, ax = ox.plot_graph(G, node_size=0)
```

[16]:
```
For the output see Figure 5
```

OSMnx also accepts place queries as unambiguous Python dictionaries to help the geocoder find a specific matching study site when several names might approximately overlap. In this example, we download the street network of San Francisco, California by defining the query with such a dictionary:

[17]:
```
# create a graph of San Francisco's drivable street network then plot it
place = {'city'   : 'San Francisco',
         'state'  : 'California',
         'country': 'USA'}
G = ox.graph_from_place(place, network_type='drive')
```

4.5 Downloading Other Infrastructure Types

All of the preceding examples have focused on urban and suburban street networks. However, OSMnx can also download and model other networked infrastructure types by passing in custom queries via the `infrastructure` argument. Such networked infrastructure could include power lines, the canal systems of Venice or Amsterdam, or the New York City subway's rail infrastructure as illustrated in this example:

[18]:
```
# create a graph of NYC's subway rail infrastructure then plot it
G = ox.graph_from_place('New York City, New York, USA',
            retain_all=False, truncate_by_edge=True, simplify=True,
            network_type='none', infrastructure='way["railway"~"subway"]')

fig, ax = ox.plot_graph(G, node_size=0)
```

[18]:
```
For the output see Figure 6
```

Figure 5: Output from codebox 16

Note that the preceding code snippet modeled subway rail *infrastructure* which thus includes crossovers, sidings, spurs, yards, and the like. For a station-based train network model, the analyst would be best-served downloading and modeling a station adjacency matrix.

Beyond networked infrastructure, OSMnx can also work with OpenStreetMap building footprint and points of interest data. For example, we can download and visualize the building footprints near New York's Empire State Building:

[19]:
```
# download and visualize the building footprints around the empire state bldg
point = (40.748482, -73.985402) #empire state bldg coordinates
dist = 812 #meters
gdf = ox.footprints_from_point(point=point, distance=dist)
gdf_proj = ox.project_gdf(gdf)
bbox_proj = ox.bbox_from_point(point=point, distance=dist, project_utm=True)
fig, ax = ox.plot_footprints(gdf_proj, bbox=bbox_proj, bgcolor='#333333',
        color='w', figsize=(6,6))
```

[19]:
```
For the output see Figure 7
```

Finally, we can download and inspect the amenities matching the tag "restaurants" near the Empire State Building and then display the five most common cuisine types

[20]:
```
# download restaurants near the empire state bldg then display them
gdf = ox.pois_from_point(point=point, distance=dist, amenities=['restaurant'])
gdf[['name', 'cuisine']].dropna().head()
```

[20]:
```
                              name        cuisine
357620442 Dolcino Trattoria Toscana        italian
419359995              Little Alley        chinese
419367625              Ramen Takumi japanese;ramen
561042187                Les Halles         french
663104998           Tick Tock Diner          diner
```

[21]:
```
# show the five most common cuisine types among these restaurants
gdf['cuisine'].value_counts().head()
```

[21]:
```
indian      22
korean      15
italian     14
japanese    13
pizza        9
Name: cuisine, dtype: int64
```

Figure 6: Output from codebox 18

Figure 7: Output from codebox 19

5 Conclusion

This article argued that computational notebooks underpin an important emerging pillar in urban analytics and planning research, pedagogy, and practice. To demonstrate this, it presented the official repository of computational notebooks that the OSMnx project uses for tutorials, demos, and guides. It illustrated the use of these notebooks by highlighting specific examples from them, inline and interactively within this article, as an introduction to this modeling and analysis software. OSMnx itself is a Python package for downloading, modeling, analyzing, and visualizing data from OpenStreetMap. It lets users analyze networked infrastructure like street networks as well as building footprints, points of interest, elevation data, and more. This article demonstrated how computational notebooks can provide a tutorial-style introduction to scientific software such as this.

The OSMnx project uses computational notebooks because they offer several advantages. First, they empower scientific reproducibility, replication, sharing, and remixing. Second, they allow researchers to intermingle data analyses with visualizations and narratives to ask and answer research questions. Third, they offer "follow-along" guides for introducing software and methods to new users, such as in this repository for OSMnx or even in the university classroom. Finally, they help researchers reach a wider community

of interest by making their methodologies and analyses more legible to a broad audience potentially interested in adapting and remixing their work. For these reasons and more, we expect to see growing adoption of computational notebooks in the urban planning discipline and related analytics fields.

References

Batty M (2019) Urban analytics defined. *Environment and Planning B: Urban Analytics and City Science* 46[3]: 403–405. CrossRef.

Boeing G (2017) Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 65: 126–139. CrossRef.

Dijkstra E (1959) A note on two problems in connexion with graphs. *Numerische Mathematik* 1[1]: 269–271. CrossRef.

Hagberg A, Schult D, Swart P (2008) Exploring network structure, dynamics, and function using networkx. In: Varoquaux G, Vaught T, Millman J (eds), *Proceedings of the 7th Python in Science Conference*. SciPy, Pasadena, CA, 11–16

Harris R, O'Sullivan D, Gahegan M, Charlton M, Comber L, Longley P, Brunsdon C, Malleson N, Heppenstall A, Singleton A, Arribas-Bel D, Evans A (2017) More bark than bytes? reflections on 21+ years of geocomputation. *Environment and Planning B: Urban Analytics and City Science* 44[4]: 598–617. CrossRef.

Kang W, Oshan T, Wolf L, Boeing G, Frias-Martinez V, Gao S, Poorthuis A, Xu W (2019) A roundtable discussion: Defining urban data science. *Environment and Planning B: Urban Analytics and City Science* 46[9]: 1756–1768. CrossRef.

Kedron P, Frazier A, Trgovac A, Nelson T, Fotheringham A (2019) Reproducibility and replicability in geographical analysis. *Geographical Analysis*. CrossRef.

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Jupyter Development Team (2016) Jupyter notebooks: A publishing format for reproducible computational workflows. In: Loizides F, Schmidt B (eds), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, Amsterdam, The Netherlands, 87–90. CrossRef.

Knuth D (1992) *Literate Programming*. Center for the Study of Language and Information, Stanford, CA

Kontokosta C (2018) Urban informatics in the science and practice of planning. *Journal of Planning Education and Research*. CrossRef.

Padgham M, Boeing G, Cooley D, Tierney N, Sumner M, Phan T, Beare R (2019) An introduction to software tools, data, and services for geospatial analysis of stroke services. *Frontiers in Neurology* 10[743]. CrossRef.

Pérez F, Granger B (2007) Ipython: A system for interactive scientific computing. *Computing in Science & Engineering* 9[3]: 21–29. CrossRef.

Poorthuis A, Zook M (2019) Being smarter about space: Drawing lessons from spatial science. *Annals of the American Association of Geographers*. CrossRef.

Porta S, Crucitti P, Latora V (2006) The network analysis of urban streets: A dual approach. *Physica A: Statistical Mechanics and Its Applications* 369[2]: 853–866. CrossRef.

Rey S (2019) Pysal: The first 10 years. *Spatial Economic Analysis* 14[3]: 273–282. CrossRef.

Singleton A, Arribas-Bel D (2019) Geographic data science. *Geographical Analysis*. CrossRef.

A Appendix

The interested reader may consult the following web sites for more information and resources as discussed in the article:

- OSMnx examples repository: https://github.com/gboeing/osmnx-examples
- OSMnx documentation: https://osmnx.readthedocs.io/
- Docker Desktop is available at: https://www.docker.com/products/docker-desktop
- The OSMnx Docker image is available at: https://hub.docker.com/r/gboeing/osmnx

REGION
The Journal of ERSA
Powered by WU

Volume 6, Number 3, 2019, 53–69
DOI: 10.18335/region.v6i3.277

journal homepage: region.ersa.org

Exploring long-term youth unemployment in Europe using sequence analysis: A reproducible notebook approach*

Nikos Patias[1]

[1] University of Liverpool, Liverpool, UK

Received: 28 August 2019/Accepted: 17 January 2020

Abstract. Youth unemployment is an important factor influencing the lifetime earnings and future job prospects of individuals, often resulting in deterioration in their health and well-being. Youth unemployment in Europe has been affected by the financial crisis of 2008. However, the magnitude of these effects varied across European countries. The objective of this notebook is to identify representative trajectories of youth unemployment change in Europe from 2008 to 2018. This notebook provides a self-contained research workflow that is fully reproducible and transparent. My findings suggest that northern Europe has high concentration of regions with stable low youth unemployment while southern Europe has high concentration of regions with stable high youth unemployment. Identifying key patterns of youth unemployment change among European countries can provide useful insights that help to understand migration patterns originating from the more "disadvantaged" regions to more "advantaged" ones, or beyond. Finally, I hope that data and regional scientists can benefit by the functionalities offered in this notebook and use it as a complementary guide for analysing their own data.

Key words: sequence analysis, unemployment, Europe, regional inequalities, reproducible research

1 Introduction

Youth unemployment is an important factor influencing the lifetime earnings and future job prospects of individuals, often resulting in deterioration in their health and well-being (Bell, Blanchflower 2011, O'Reilly et al. 2015). The effects of financial crisis of 2008 on youth unemployment were prominent and varied across European countries and regions. The European average for youth unemployment culminated in 2012 to more than 20%, but there were countries that scored much higher (i.e. more than 50% in Greece and more that 30% in Bulgaria and Italy) (Dietrich 2012). However, regional variations can help to contextualise and analyse patterns of youth unemployment more effectively (Pop et al. 2019). Understanding and tracking the evolution of regions that faced high levels of unemployment can help in planning future policies, as today's youth are going to be in the workforce for the next 50 years. Finally, the trajectories of youth unemployment change across regions (i.e. whether they have successfully recovered or not) and can be

*This paper is available as computational notebook on the REGION webpage.

linked to patterns of regional resilience against economic crises. In this notebook, I use a sequence analysis approach to identify representative trajectories of youth unemployment change by NUTS 2 regions in Europe from 2008 to 2018.

The idea of using reproducible analyses in computational research has a growing number of advocates (Peng 2011, Sandve et al. 2013, Rule et al. 2019). The development of computational notebooks such as R and Jupyter notebooks, allow scientists to incorporate code, documentation, graphs and text in a single document. Consequently, more than ever before, computational research has become more open, transparent and fully replicable. Peng (2011) developed a reproducibility spectrum to highlight the importance of incorporating publication standards text with linked code and data to achieve the "gold standard of reproducibility". The spectrum begins from the traditional static publications, which are not reproducible. They become more reproducible when code and data are incorporated in the publication. Finally, the full replication is achieved when linked and executable code and data are included. As Peng (2011) highlights, data is an integral part of reproducible research and should be clearly documented within the workflow. However, researchers often neglect to provide adequate information on the datasets used. As a result, other researchers have difficulties on replicating this piece of research. Linked datasets through direct web-links or Application Programming Interfaces (APIs) when available, contribute to the transparency of the research, by explicitly pointing the end-user to the source of information described in a research project.

The objective of this notebook is to identify key representative trajectories of youth unemployment change in Europe from 2008 to 2018. In the present notebook I provide a self-contained research workflow that is fully reproducible and transparent. Moreover, I make use of the functionalities offered by computational notebooks written in R markdown such as direct access to online tabular/spatial datasets, manipulation and linkage between these datasets as well as interactive plots/maps. Finally, this notebook aims to provide the sufficient tools that a data or regional scientist needs to perform similar types of analysis.

2 Packages and Dependencies

This section is used to report all the packages and dependencies required to run this notebook which are vital components of reproducible research. By reporting the R version under which I created this notebook and the packages used will ensure its replicability.

Firstly, is important to report the R version used in this notebook by running the following line of code.

[1]:
```
# to get the version of R used in the notebook
paste("The R Version used in this notebook is", getRversion())
```

[1]:
```
## [1] "The R Version used in this notebook is 3.5.1"
```

I then specify the CRAN repository where the packages have been downloaded from.

[2]:
```
# Define the CRAN repository for this session
r_rep = getOption("repos")
r_rep["CRAN"] = "http://cran.us.r-project.org"
options(repos = r_rep)
```

And install/load the packages required to run this notebook. Please note that the installation stage is required only the first time you run this notebook.

[3]:
```
# These are the packages required to run this notebook
# First should be installed
# install.packages("eurostat")
# install.packages("rvest")
# install.packages("knitr")
# install.packages("rgdal")
# install.packages("countrycode")
# install.packages("dplyr")
# install.packages("reshape2")
# install.packages("ggplot2")
# install.packages("TraMineR")
# install.packages("cluster")
```

```
# install.packages("factoextra")
# install.packages("RColorBrewer")
# install.packages("leaflet")
# install.packages("plotly")
# And then should be loaded
library(eurostat)
library(rvest)
library(knitr)
library(rgdal)
library(countrycode)
library(dplyr)
library(reshape2)
library(ggplot2)
library(TraMineR)
library(cluster)
library(factoextra)
library(RColorBrewer)
library(leaflet)
library(plotly)
```

Finally, I create a list of the available packages in my R environment and report the version of each package used.

[4]:
```
# Create a list with all the available packages in my R environment
pkg_used <- available.packages()
```

[5]:
```
# For every package print the version of the package, the version of R that depends
# on and the packages that imports
paste("eurostat Version is:", pkg_used["eurostat", "Version"])
paste("rvest Version is:", pkg_used["rvest", "Version"])
paste("knitr Version is:", pkg_used["knitr", "Version"])
paste("rgdal Version is:", pkg_used["rgdal", "Version"])
paste("countrycode Version is:", pkg_used["countrycode", "Version"])
paste("dplyr Version is:", pkg_used["dplyr", "Version"])
paste("reshape2 Version is:", pkg_used["reshape2", "Version"])
paste("ggplot2 Version is:", pkg_used["ggplot2", "Version"])
paste("TraMineR Version is:", pkg_used["TraMineR", "Version"])
paste("cluster Version is:", pkg_used["cluster", "Version"])
paste("factoextra Version is:", pkg_used["factoextra", "Version"])
paste("RColorBrewer Version is:", pkg_used["RColorBrewer", "Version"])
paste("leaflet Version is:", pkg_used["leaflet", "Version"])
paste("plotly Version is:", pkg_used["plotly", "Version"])
```

[5]:
```
## [1] "eurostat Version is: 3.4.20002"
## [1] "rvest Version is: 0.3.5"
## [1] "knitr Version is: 1.27"
## [1] "rgdal Version is: 1.4-8"
## [1] "countrycode Version is: 1.1.0"
## [1] "dplyr Version is: 0.8.3"
## [1] "reshape2 Version is: 1.4.3"
## [1] "ggplot2 Version is: 3.2.1"
## [1] "TraMineR Version is: 2.0-14"
## [1] "cluster Version is: 2.1.0"
## [1] "factoextra Version is: 1.0.6"
## [1] "RColorBrewer Version is: 1.1-2"
## [1] "leaflet Version is: 2.0.3"
## [1] "plotly Version is: 4.9.1"
```

In this notebook, I have installed the latest versions of the packages used. I understand that the analysis can be run by using previous versions too. However, using the versions of the packages as reported here ensures the reader that this notebook will run without any errors.

Now that all the packages have been correctly installed it is useful to provide a brief overview of the main functionalities of each package.

`eurostat` This package allows access to Eurostat data through their API.

`rvest` This package is used to scrape data from web pages.

`knitr` This package provides better visualisation of the results within the notebook (i.e. table formatting).

`rgdal` This package is used to read, merge and manipulate geospatial datasets.

`countrycode` This package is used to convert country codes (i.e. ISO 3166) to country names.

`dplyr` This package is used for more effective dataframes' manipulation.

`reshape2` This package is used to reshape tables from wide to long format and vice versa.

`ggplot2` This package is used for creating plots.

`TraMineR` This is the package is used to perform sequence analysis.

`cluster` This package is used to perform cluster analysis.

`factoextra` This package provides the functionalities to assess the optimal number of clusters.

`RColorBrewer` This package provides colour palettes to be used in maps.

`leaflet` This package is used for interactive mapping.

`plotly` This package is used for interactive plotting in conjuction with `ggplot2`.

3 Data and Methods

3.1 Data

Eurostat (https://ec.europa.eu/eurostat/data/database) has a large database providing a wide range of available datasets in varying geographies and time frames. In this notebook, I analyse youth unemployment in Europe from 2008 to 2018. Eurostat captures youth unemployment by measuring the percentage of "Young people neither in employment nor in education and training (NEET rates)". This dataset is available from 2000 to 2018 at NUTS 2 regions (more information on NUTS classification can be found at https://ec.europa.eu/eurostat/web/nuts/background). Eurostat defines youth unemployment either people aged 15-24 or 18-24 and are unemployed. In this study, I consider the percentage of people aged between 18 and 24. The original dataset used in this notebook can be accessed following https://ec.europa.eu/eurostat/web/products-datasets/-/edat_lfse_22. Eurostat has created its own R package (https://cran.r-project.org/web/packages/eurostat/index.html) to allow access in the database through an API with a comprehensive documentation (https://ropengov.github.io/eurostat/index.html) and tutorial (http://ropengov.github.io/eurostat/articles/eurostat_tutorial.html). In order to make the most of the functionalities offered by a notebook as well as enable researchers to replicate this approach I make use of Eurostat's API to access the dataset analysed in this notebook.

I first search for the datasets referring to young unemployed people.

[6]:
```
# search information about the datasets that are related to young unemployed people
kable(head(search_eurostat("Young people neither in employment")))
```

[6]:
```
Output in Table 1
```

Of the available datasets, I am interested in the first on this list with code `edat_lfse_22`. I specified my request to 11 time periods. Starting from the most current year (i.e. 2018) and going back to 2008. I also specified that I want total percentages (i.e. both male and female - `sex = "T"`) and finally the age group that covers people aged from 18 to 24.

[7]:
```
# Specify the ID of the dataset required
id <- "edat_lfse_22"
# Request of the dataset
young_unempl <- get_eurostat(id, filters = list(lastTimePeriod=11, sex = "T",
                     age = "Y18-24"), time_format = "num")
```

Table 1: Available datasets from Eurostat related to youth unemployment

title	code	type	last update of data	last table structure change	data start	data end	values
Young people neither in employment nor in education and training by sex and NUTS 2 regions (NEET rates)	edat_lfse_22	dataset	01.07.2019	13.08.2019	2000	2018	NA
Young people neither in employment nor in education and training by sex, age and degree of urbanisation (NEET rates)	edat_lfse_29	dataset	01.07.2019	13.08.2019	2000	2018	NA
Young people neither in employment nor in education and training by type of disability, sex and age	hlth_de030	dataset	21.03.2019	21.03.2019	2011	2011	NA
Young people neither in employment nor in education and training by sex, age and labour status (NEET rates)	edat_lfse_20	dataset	01.07.2019	13.08.2019	2000	2018	NA
Young people neither in employment nor in education and training by sex, age and citizenship (NEET rates)	edat_lfse_23	dataset	25.04.2019	13.08.2019	2004	2018	NA
Young people neither in employment nor in education and training by sex, age and country of birth (NEET rates)	edat_lfse_28	dataset	25.04.2019	13.08.2019	2004	2018	NA

I also make use of a spatial dataset to enable use of an interactive map to facilitate better presentation of results. To achieve that, I have downloaded NUTS 2 regions shapefile from the second version of Eurostat's spatial database (https://ec.europa.eu/eurostat/cache/GISCO/distribution/v2/). Eurostat provides the spatial data as a bulk download, so I first unzip the file and then select the shapefile that matches the tabular data that I have already downloaded. The name of the dataset "NUTS_RG_60M_2016_4326_LEVL_2.shp.zip" is self-explanatory showing that the spatial data is projected in the World Geodetic System of 1984 (i.e. WGS84 or EPSG:4326) for NUTS 2 regions in 2016.

[8]:
```
# Specify the url that links to the zipped spatial datasets
url = "http://ec.europa.eu/eurostat/cache/GISCO/distribution/v2/nuts/download/
...ref-nuts-2016-60m.shp.zip"
# Download the file
download.file(url, basename(url))
# Unzip the bulk file
unzip(basename(url))
# Unzip the specific shapefile needed
unzip(paste0(getwd(), "/NUTS_RG_60M_2016_4326_LEVL_2.shp.zip"))
# Read in the shapefile
geodata <- readOGR(dsn = getwd(), layer = "NUTS_RG_60M_2016_4326_LEVL_2")
```

Eurostat provides only country codes, which is not always helpful when presenting results. For this reason, I used the R package `countrycode` (https://cran.r-project.org/web/packages/countrycode/index.html) to convert country codes to country names. While in general, the European commission uses ISO 3166-1 alpha-2 codes, there are two exceptions. Greece is reported as "EL" (rather than "GR") and United Kingdom as "UK" (rather than "GB"). Thus, I recoded these two countries manually.

[9]:
```
# Create a new column for country names
geodata@data$cntr_name <- countrycode(geodata@data$CNTR_CODE, "iso2c", "country.name")
# Because European commission uses EL for Greece (in ISO 3166-1 alpha-2 codes is GR)
# and UK for United Kingdom (in ISO 3166-1 alpha-2 codes is GB) I should replace these
# two countries manually
geodata@data$cntr_name <- ifelse(geodata@data$CNTR_CODE=="EL","Greece",
ifelse(geodata@data$CNTR_CODE=="UK", "United Kingdom", geodata@data$cntr_name))
```

3.2 Methods

This notebook aims to identify representative trajectories of youth unemployment change across NUTS 2 regions in Europe. The methodological workflow followed here is similar to Patias et al. (2020) that analyses trajectories of neighbourhood change in Great Britain. Sequence analysis is a method that analyses sequences of categorical variables, and extracts information on their structure and evolution. Sequence analysis has its origins in biology, where it is used to analyse DNA sequences (Sanger et al. 1977). It can also be applied to analyse longitudinal individual-level family, migration and career trajectories (Brzinsky-Fay 2007, Rowe et al. 2017a,b). This method is also used on neighbourhood trajectory mining in the United States to identify patterns of socioeconomic change over

a period of time (Delmelle 2016). The key component of sequence analysis method is the optimal matching analysis which is used to measure pairwise dissimilarities between sequences and identifies "types of sequence patterns" (Studer, Ritschard 2016). In this notebook sequence analysis is used in a spatio-temporal concept assessing how youth unemployment in European regions (i.e. spatial) has changed from 2008 to 2018 (i.e. temporal). Sequence analysis has been used as it is a method capable of capturing multiple dimensions of spatio-temporal processes namely incidence, duration, timing and sequencing. The youth unemployment data downloaded from Eurostat is expressed in percentages (by NUTS 2 regions). Sequence analysis can be applied to categorical data, hence I had to classify the regions' youth unemployment percentages into quintiles so that they can be treated as categories. In this way, the multiple dimensions of spatio-temporal processes of youth unemployment change can be systematically measured. Thus, it explicitly captures for each region:

- The number of times in a particular quintile (i.e. incidence);
- The time span in a particular quintile (i.e. duration);
- The year at occurrence of change from one quintile to another (i.e. timing); and
- The chronological order of transitions between quintiles (i.e. sequencing).

The key stages followed in this notebook are described below with links to the particular sub-sections which provide some further technical clarifications:

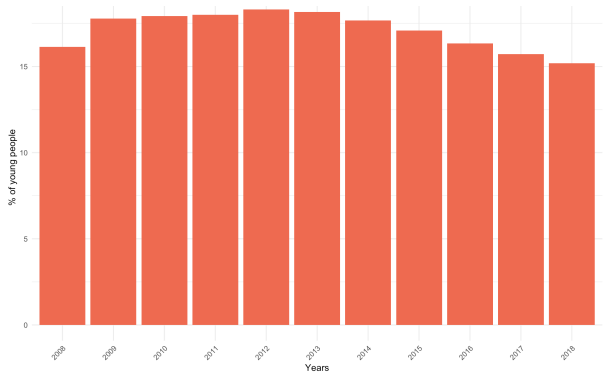
1. *Data pre-processing*, by classifying NUTS 2 regions into quintiles based on the percentage of youth unemployment in each year from 2008 to 2018 (regions with the lowest % youth unemployment belong to the 1st quintile and regions with the highest % youth unemployment belong to the 5th quintile).
2. Create a *sequence object* based on the quintile each region belongs to in every year (i.e. from 2008 to 2018).
3. *Measuring sequence dissimilarity* based on substitution costs which is the probability of transitioning from one quintile to another (i.e. higher transition rate from 1st quintile to 5th quintile rather than from 2nd quintile to 1st quintile). The substitution costs between quintiles i and j are calculated based on Equation (1).
4. Using the substitution costs calculated in the previous stage, I built a *dissimilarity matrix* including every pair of sequences. In this notebook I have used the Optimal Matching (OM) algorithm. The algorithm substitutes the elements of each sequence based on their substitution costs which in turn is the OM distance between each pair of sequences.
5. In the last stage I produce I typology of youth unemployment trajectories using the resulting dissimilarity matrix from stage 4. Partitioning Around Medoids (PAM) clustering algorithm is used for the *classification of sequences*

$$SubsCosts_{i,j} = 2 - p(i|j) - p(j|i) \tag{1}$$

where $p(i|j)$ is the transition rate between quintiles i and j. For the sequence analysis I have used R package `TraMineR` (https://cran.r-project.org/web/packages/TraMineR/index.html) which provides all the required functionalities.

4 Data Analysis

As shown in Figure 1, the average percentage of young people who are neither in employment nor in education and training in Europe increased from 16% in 2008 to 18.5% in 2012, followed by a decrease in 2018 (i.e. at around 15%). The results suggest that on average, regions show patterns of resilience against financial crises and that policies



Notes: Data from Eurostat, calculations by the author

Figure 1: European % average of young people neither in employment nor in education and training

targeting the decrease of youth unemployment have proven efficient. However, not all regions follow the same patterns.

This section of the notebook aims to provide an understanding on long-term youth unemployment patterns in NUTS 2 regions in Europe but also to guide the reader on the analytical process of sequence analysis and the functionalities offered by computational notebooks. Each of the following five sub-sections will present in detail each of the steps followed for the production of the results.

[10]:
```
# I change the years from numbers to characters so to be recongised as categorical
# rather than continuous variable
young_unempl$time <- as.character(young_unempl$time)
# Create a plot by showing the European % average of young people neither in employment
# nor in education and training
young_unempl %>%
  group_by(time) %>%
  summarise_all(mean, na.rm = TRUE) %>%
  ggplot()
  geom_bar(aes(x = time, y = values), stat = "identity", fill = "coral2")
  labs(title = "European average % of young unemployed people",
     x = "Years",
     y = "% of young people",
     caption = "Data downloaded from Eurostat\ncalculations made by the author")
  theme_minimal()
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

[10]:
```
Output in Figure 1
```

4.1 Data pre-processing

While the datasets provided by Eurostat are in clean format, they almost always require some data pre-processing. Here, there are three main tasks required to bring the data in the required format to proceed with the analysis. First, to subset the dataset to have only the NUTS 2 regions (the original dataset also includes country and NUTS 1 data). Second, to calculate the quintiles that every NUTS 2 region belongs to in every year. Third, to re-format the dataset from a long (each region represented in multiple rows – one for every year) to a wide format (each region represented by a single row and there are multiple columns containing the yearly young unemployment rates).

In coding terms, the first task is to calculate the number of characters of all geography codes and store them in a new column. I then create a subset of the dataset with only NUTS 2 regions which are those that contain four characters (i.e. the first two represent the country name followed by two numbers represent the NUTS 2 region).

Table 2: Preview of the data that will be used in sequence analysis

	sex	age	training	wstatus	unit	geo	n_char	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
2	T	Y18-24	NO_FE_NO_NFE	NEMP	PC	AT12	4	2	2	1	2	1	1	2	1	1	1	1
3	T	Y18-24	NO_FE_NO_NFE	NEMP	PC	AT13	4	2	2	2	2	2	2	2	3	3	2	3
5	T	Y18-24	NO_FE_NO_NFE	NEMP	PC	AT22	4	1	2	1	1	1	1	1	1	1	1	1
6	T	Y18-24	NO_FE_NO_NFE	NEMP	PC	AT31	4	1	1	1	1	1	1	1	1	1	1	1
8	T	Y18-24	NO_FE_NO_NFE	NEMP	PC	AT33	4	1	1	1	1	1	1	1	1	1	1	1
10	T	Y18-24	NO_FE_NO_NFE	NEMP	PC	BE10	4	5	4	4	4	4	4	4	4	4	4	4

[11]:
```
# Create a new column to store the number of characters of the geography
young_unempl$n_char <- nchar(as.character(young_unempl$geo))
# Subset only the NUTS 2 regions - their geography code contains 4 characters
young_unempl_NUTS2 <- young_unempl %>%
  filter(n_char == 4)
```

The next task is to calculate quintiles by year for each region based on their % of youth unemployment. This was done by looping through each year.

[12]:
```
# Calculate quintiles by year
# It is good to specify  the filter function to be used from dplyr function to avoid
# error messages
quant_data <- NULL
for (var in unique(young_unempl_NUTS2$time)) {
    young_unempl_NUTS2_temp <- young_unempl_NUTS2 %>%
      dplyr::filter(time == var) %>%
      mutate(quintiles = ntile(values, 5) )
    quant_data <- rbind(quant_data,young_unempl_NUTS2_temp)
}
```

The final task is to keep only the column containing the quintiles (the actual percentages will not be used in the rest of this notebook). The dataset will then be re-formatted to a wide format where each region will be represented by a single row. For each region there are multiple columns containing the corresponding yearly young unemployment quintiles. Finally, I delete all the rows that contain missing values. This is important as there are regions that have "gaps" in their data availability, meaning that there is no data in at least one year between 2008 and 2018. These regions are ignored in this analysis to speed up computational time and to have consistency across sequences.

[13]:
```
# I delete the column inlcuding the % as I will use the quintiles from now on in the
# analysis
quant_data <- subset(quant_data, select = -values)
# Re-format the data from long to wide format
# This means that every row will represent a region and every column represents a year
quant_data_wide <-
  dcast(quant_data,
  sex  age  training  wstatus  unit  geo  n_char ~ time,
  value.var = 'quintiles')
# We remove rows that  do not have values in at least one year so we have consistency
# between sequences
quant_data_wide <- na.omit(quant_data_wide)
# Have a look at the dataset
kable(head(quant_data_wide))
```

[5]:
```
Output in Table 2
```

4.2 Sequence object

Creating a sequence object is the initial point of sequence analysis. In this notebook I pass a subset of the columns of the dataset that contain the quintile values. Practically, it means that I create a sequence of quintiles from 2008 to 2018 which are from the 8th to 18th column for each region. As I have already mentioned, I have used the R package `TraMineR` (https://cran.r-project.org/web/packages/TraMineR/index.html). For more detailed information on sequence analysis and all the functionalities, please refer to the user guide (http://mephisto.unige.ch/pub/TraMineR/doc/TraMineR-Users-Guide.pdf) which provides detailed information on all the functionalities of the package.

Table 3: Substitution costs between quintiles

	1	2	3	4	5
1	0.000000	1.694604	1.985533	2.000000	2.000000
2	1.694604	0.000000	1.591372	1.960797	2.000000
3	1.985533	1.591372	0.000000	1.654430	2.000000
4	2.000000	1.960797	1.654430	0.000000	1.752492
5	2.000000	2.000000	2.000000	1.752492	0.000000

[14]:
```
# Create the sequence object using only the quintiles that every region belongs
seq_obj <- seqdef(quant_data_wide[,8:18])
```

4.3 Measuring sequence dissimilarity

A key element of sequence analysis is to calculate "distances" between each pair of sequences that can be used later for the Optimal Matching analysis. These distances are a measure based on how similar two sequences are. There are two components related to these distances. First is the insertion/deletion (indel) cost which is used when the length of sequences is not the same. This is the cost of deleting or inserting a state in a sequence so all the sequences have the same length. On this notebook, the time period covered is the same for every region (i.e. from 2008 to 2018) so the sequence length is fixed, an 11-state long sequence. Hence this step is not required.

The second component for calculating "distances" between each pair of sequences is to calculate the substitution costs for transforming one state (i.e. one quintile group) to another. Substitution costs can be theory-driven or empirically-driven (Salmela-Aro et al. 2011). Theory-driven costs are usually used when researchers define costs based on pre-determined concepts. Thus, the costs between states are solely dependent on the researchers' choices (i.e. how "far" is one state from another). On the other hand, empirically-driven costs are based on the observed transitions between states. Hence, two states are closer when there are more observed transitions between them. In this notebook I follow the empirically-driven approach as I intend to explicitly consider the observed transitions between states (i.e. quintile groups here).

By following the empirically-driven approach, there are two options to calculate substitution costs. The first is to assign a constant value for substituting sequence states (i.e. quintiles). The second option is to calculate transition rates which are the probabilities of transitioning from one state to another (between quintiles in this notebook). These transition rates are then used to calculate the substitution costs as shown in Equation (1). In this analysis, I have used transition rates (i.e. `method = "TRATE"`) because it is important to capture the higher probability of transitioning between 1st and 2nd quintile compared to 1st and 5th quintile. By assigning a constant value, this information would have been missed. For more detailed information on substitution costs please refer again to the `TraMineR` user guide (http://mephisto.unige.ch/pub/TraMineR/doc/TraMineR-Users-Guide.pdf).

Table 3 shows the substitution costs of this study. It is clear from the table that the probability of transitioning between 1st and 2nd quintile is higher than transitioning from 1st to 5th quintile. Hence, the substitution cost from 1st to 2nd is lower than the 1st to 5th. This information will then be used in the next step – the Optimal Matching.

[15]:
```
# Calculate substistution costs
subs_costs <- seqsubm(seq_obj, method = "TRATE")
# Print the substitution costs
kable(subs_costs)
```

[5]:
```
Output in Table 3
```

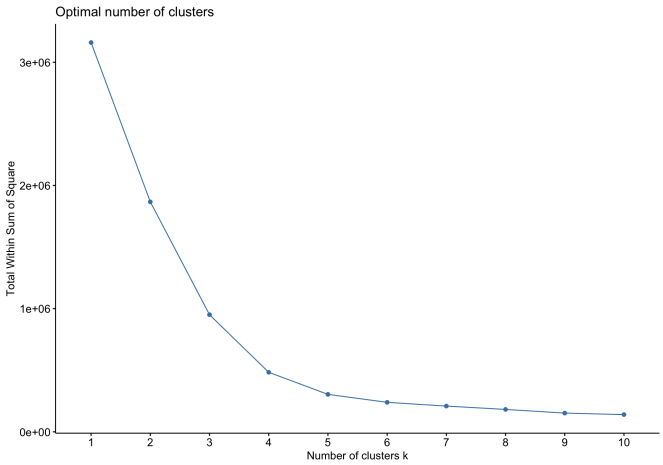


Figure 2: Within sum of squares to access optimal clustering solution

4.4 Dissimilarity matrix

The dissimilarity matrix is a symmetric matrix between all the regions. The matrix is populated by calculating the difference of the sequences between every pair of regions and is symmetrical because each row and column represents a NUTS 2 region (as it happens in any distance matrix). Each region consists of an 11-state long sequence (i.e. from 2008 to 2018). As in the previous stages there are different options to compare sequences. In this notebook I have used the simple Optimal Matching algorithm which uses the substitution costs between the quintiles and aggregates them for every pair of sequences. 'Dynamic Hamming' distance is an alternative method of Optimal Matching which calculates different substitution costs for every time period, assuming that the probabilites of transitioning between different states significantly change over time. Another variant of Optimal Matching is the 'Optimal Matching of Transition Sequences' that accounts for the sequencing of states by explicitly considering their order. In this notebook I have used the simple Optimal Matching algorithm as the aim of the notebook is to present how sequence analysis can be applied to the context of exploring youth unemployment change without making any assumptions that the timing or the ordering of states is considered more important which other Optimal Matching methods can explicitly capture. Studer, Ritschard (2016) provide a good review of different variants of Optimal Matching.

Hence, using the Optimal Matching method a dissimilarity matrix between all NUTS 2 regions has been built based on the substitution costs shown in Table 3. Lower costs mean that sequences are more similar, while larger costs mean that they are different. Hence, it is an abstract distance matrix, showing how "close" two sequences are.

[16]:
```
# Calculate the distance matrix
seq.OM <- seqdist(seq_obj, method = "OM", sm = subs_costs)
```

4.5 Classification of sequences

The final analytical step is to classify the sequences based on their similarities. There is a wide range of clustering algorithms to choose from, when it comes to object classification. Here, the Partitioning Around Medoids (PAM) clustering method was selected for classifying sequences. The PAM algorithm is similar to k-means, but is considered more robust (Kaufman, Rousseuw 1991). A dissimilarity matrix can be used as an index. The algorithm iterates to minimize the sum of dissimilarities within clusters, compared to k-means that aims to minimize the sum of squared Euclidean distances. PAM is based on finding k representative objects or medoids among the observations and then k clusters (that should be defined as in k-means) are created to assign each observation to its nearest medoid. There are different fit statistics to assess optimal clustering solutions.

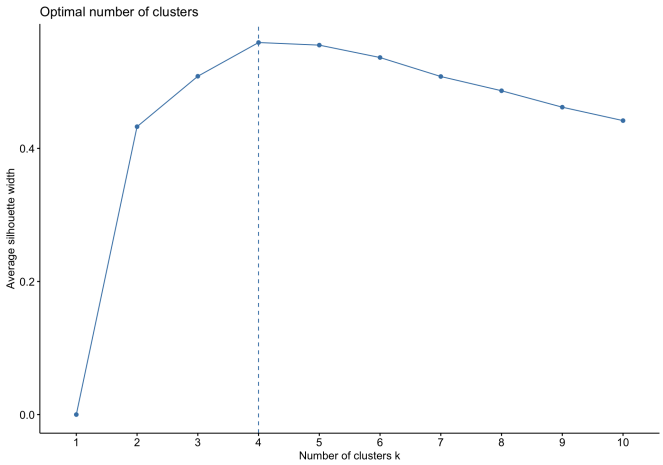


Figure 3: Average silhouette width to access optimal clustering solution

[17]:
```
# Assess different clustering solutions to specify the optimal number of clusters
fviz_nbclust(seq.OM, cluster::pam, method = "wss")
```

[17]:
```
Output in Figure 2
```

[18]:
```
# Assess different clustering solutions to specify the optimal number of clusters
fviz_nbclust(seq.OM, cluster::pam, method = "silhouette")
```

[18]:
```
Output in Figure 3
```

In this notebook I have used two fit statistics (see Figures 2 and 3) to assess various clustering solutions. The focus of this notebook is not to demonstrate the differences between fit statistics. Thus, I will not get into more detail on what every measure means. However, a useful tutorial can be found at https://rstudio-pubs-static.s3.amazonaws.-com/455393_f20bacf1329a49dab40eb393308b33eb.html. In short, they show how well separated each cluster is compared to other clusters (see Figure 2) but also how "compact" the observations are within each cluster (see Figure 3). The fit statistics here show that the optimal clustering solution is four clusters.

[19]:
```
# Run clustering algorithm with k = 4
pam.res <- pam(seq.OM, 4)
```

Having classified the sequences, it is then important to visualise the results to understand differences between the groups. Figure 4 and 5 show the four resulting transition patterns of young unemployment in NUTS 2 regions based on the quintiles they belonged from 2008 to 2018. In Figure 4 each line represents a region, each colour a quintile group and the x-axis represents each year. Figure 5 displays the year-specific distribution of each sequence group. Finally, the y-axis in Figure 4 represents the total number of sequences within each sequence group, while in Figure 5 it represents the distribution of sequences that belong to each sequence group at thus it ranges from 0 to 1.

[20]:
```
# Assign the cluster group into the tabular dataset
quant_data_wide$cluster <- pam.res$clustering
# Then rename clusters
quant_data_wide$cluster <- factor(quant_data_wide$cluster, levels=c(1, 2, 3, 4),
                                  labels=c("Stable Low youth unemployment",
                                           "Stable Moderate youth unemployment",
                                           "Increasingly High youth unemployment",
                                           "Stable High youth unemployment"))
```

For convenience and better communication of the results I assigned names to the four groups starting from the top left plot as:

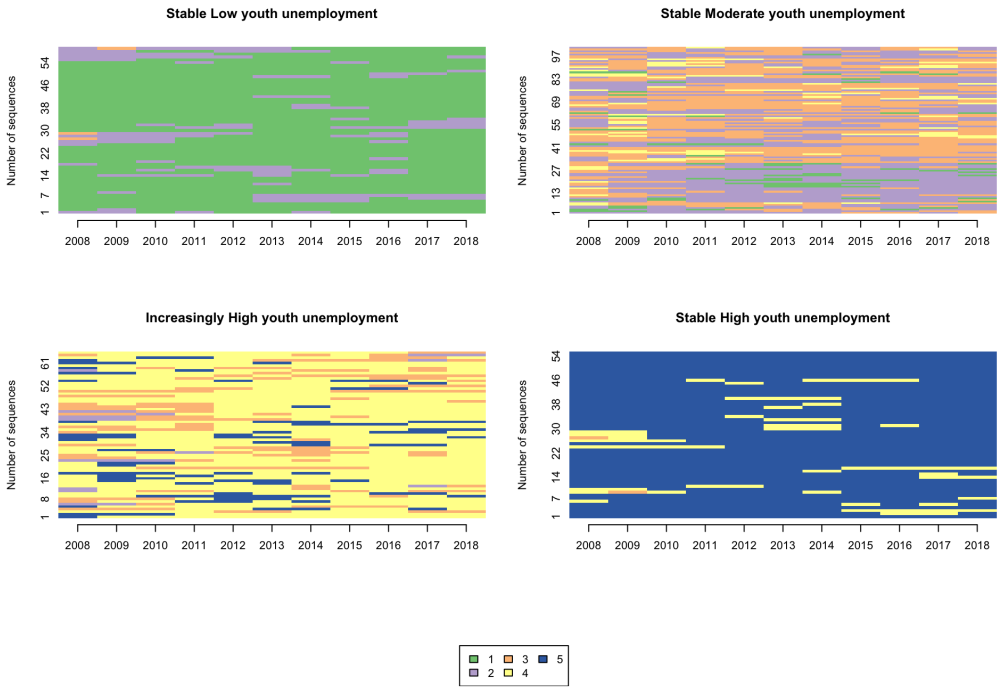


Figure 4: Individual sequences by sequence group

- Group 1 ... Stable Low youth unemployment
- Group 2 ... Stable Moderate youth unemployment
- Group 3 ... Increasingly High youth unemployment
- Group 4 ... Stable High youth unemployment

[21]:
```
# Plot of individual sequences split by sequence group
seqIplot(seq_obj, group = quant_data_wide$cluster, ylab = "Number of sequences")
```

[21]:
```
Output in Figure 4
```

[22]:
```
# Distribution plot by sequence group
seqdplot(seq_obj, group = quant_data_wide$cluster, border=NA,
ylab = "Distribution of sequences")
```

[22]:
```
Output in Figure 5
```

The results of this analysis mainly show patterns of stability in terms of youth unemployment. Group 1 contains regions that are in the lowest quintile, which means they have the lowest youth unemployment ratios over time. Group 4 is exactly the opposite of group 1, containing the regions that belong to the highest quintile over time (highest youth unemployment ratios). Group 2 consists of regions that classified either in the 2nd or 3rd quintile in the last 10 years. Finally, group 3 consists of regions that initially (i.e. 2008) belonged to 3rd, 4th, 5th and few on the 2nd quintile but gradually transformed to the 4th quintile, thus now having a higher percentage of young unemployed people.

5 Exploring spatio-temporal trends of youth unemployment in Europe

Youth unemployment as a socioeconomic phenomenon is of main concern in European policy. Thus, it is important to visualise the findings of this notebook, so that they can be

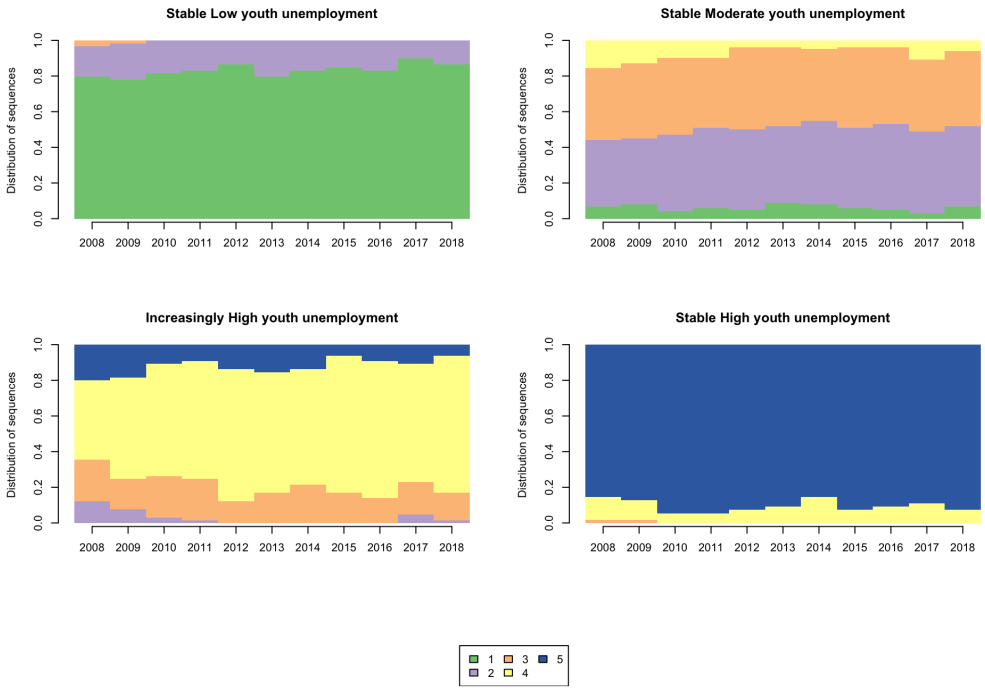


Figure 5: Distribution plot by sequence group

easily explored by the reader. To achieve this, I linked the results of the sequence analysis to the NUTS 2 region geographies so to create an interactive map. I then calculated the frequency of each trajectory group in every European country and created an interactive plot. In this way, the results are more accessible to everyone interested[1].

The map (see Figure 6) offers the opportunity to hover over the regions. Then by clicking on any region, information on the trajectory group, the region name and the country name is shown. The interactive plot (see Figure 7) offers an overview of the frequencies of trajectory groups across European countries. By hovering over the plot, one can observe the exact frequency of each group. It also offers the opportunity to zoom in on particular countries and to manually navigate through the graph (i.e. pan option on the toolbox on the right top of the plot). Finally, by clicking on the legend, particular group(s) can be selected to be shown.

[23]:
```
# Merge the spatial to the tabular dataset which includes the cluster names
map_data <- merge(geodata, quant_data_wide, by.x="FID", by.y="geo", all.x=TRUE)
```

Figures 6 and 7 show spatio-temporal variations of youth unemployment within and across European countries. As illustrated in the map, Mediterranean and Balkan countries (i.e. Greece, Italy, Spain, Turkey, Bulgaria and Romania) have stable high youth unemployment over time. On the other hand, northern countries and central European countries (i.e. Norway, Sweden, Netherlands, Germany, Austria and Switzerland) have stable low youth unemployment over time. Finally, the majority of central European countries and the United Kingdom have followed moderate levels of youth unemployment change. However, there are regional differences highlighting that socioeconomic inequalities are not only apparent between countries but also within their national boundaries. There is a clear split between the stable high youth unemployment in south Italy compared to lower but still increasingly high youth unemployment in the north. Spain has three tiers, split geographically, where the more northern the region, the lower the youth unemployment level. A similar pattern appeared in the United Kingdom, where northern regions have higher youth unemployment levels than southern regions over time.

[1]The interactive figures are included in the HTML-version of the paper or can be generated from the Rmd-file.

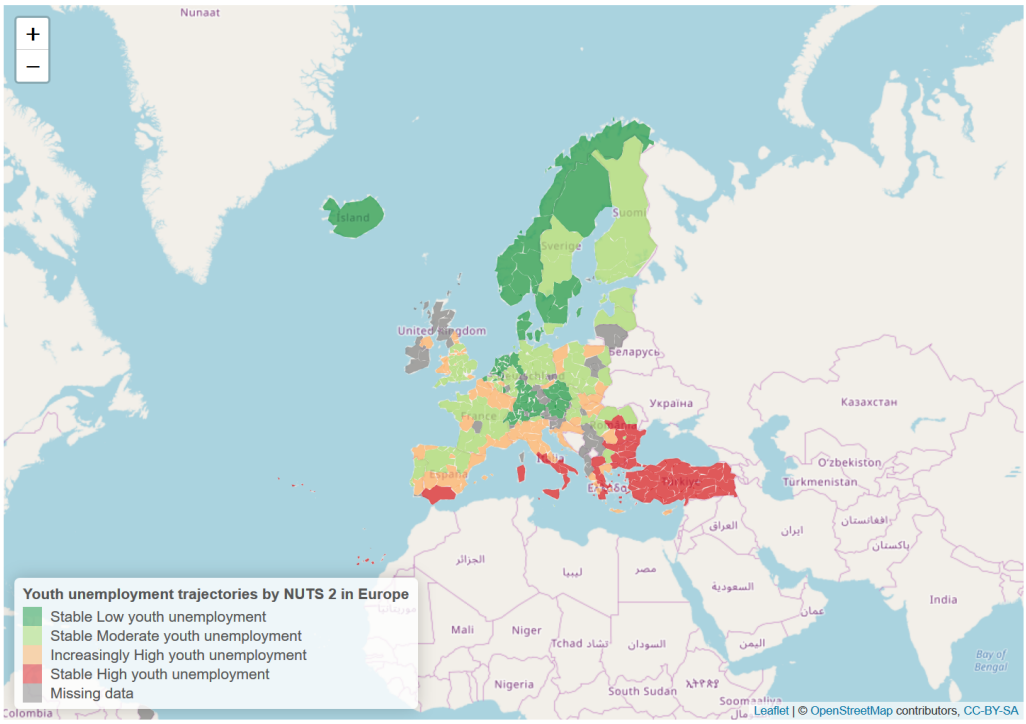


Figure 6: Interactive map of youth unemployment trajectories in NUTS 2 from 2008 to 2018

When looking in more detail at some major metropolitan regions, we can observe deviations from their neighbouring regions. Bucharest and Sofia seem to have lower unemployment levels compared to adjacent regions in Romania and Bulgaria respectively. On the other hand, while Belgium has moderate or low levels of youth unemployment on average, Brussels, its biggest city and capital, has stable higher youth unemployment. Austria follows similar pattern where the country has on average high concentration of 'stable low youth unemployment' regions but its biggest city (and capital) Vienna is classified as 'stable high youth unemployment'. This highlights that higher levels of socioeconomic inequalities and more disadvantaged groups are often aggregated in large metropolitan areas.

[24]:
```
# Create a map showing the distribution of sequence clusters
# Specify the colour palette
myColors <- rev(brewer.pal(4,"RdYlGn"))
pal <- colorFactor(myColors, domain = unique(map_data$cluster))
# Create the initial background map, zooming in Europe
colourmap <- leaflet() %>%
  addTiles() %>%
  setView(lat = 55, lng = 1, zoom = 3)
# Create the interactive map showing the sequence clusters
colourmap %>%
  addPolygons(data = map_data,
    fillColor = ~pal(cluster),
    weight = 0.2,
    opacity = 0.8,
    color = "white",
    dashArray = "3",
    fillOpacity = 0.7,
    popup = paste("Cluster: ", map_data$cluster, "<br>",
                  "NUTS 2 Name: ", map_data$NUTS_NAME, "<br>",
                  "Country Name: ", map_data$cntr_name, "<br>"),
    highlight = highlightOptions(
      weight = 5,
      color = "#666",
      dashArray = "",
```

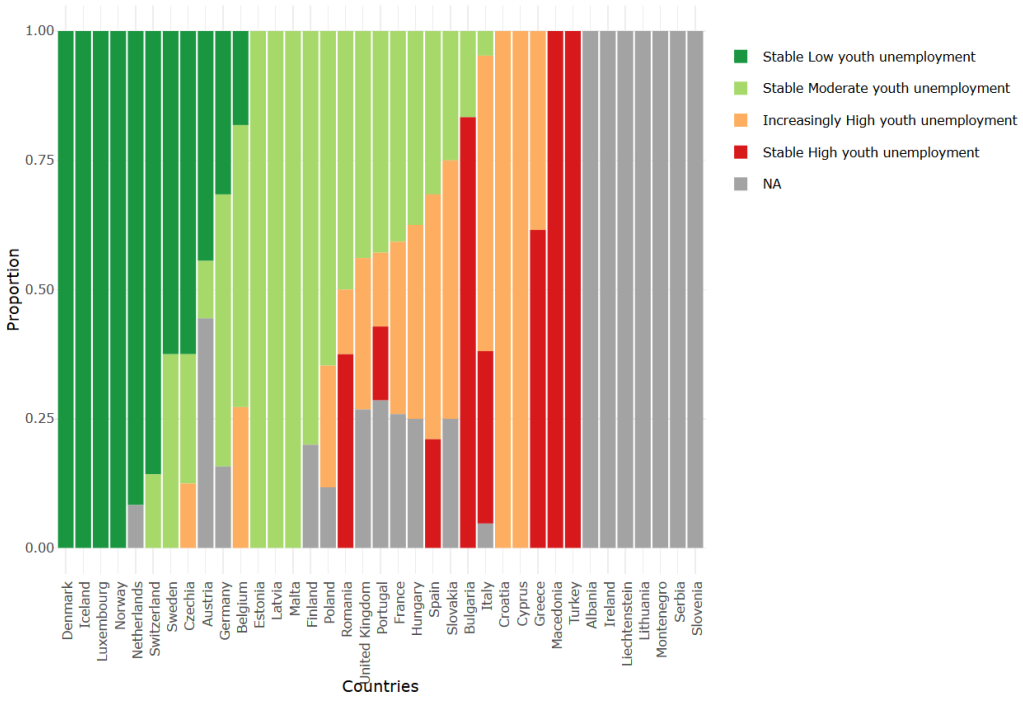


Figure 7: Distribution of youth unemployment trajectories across Europe from 2008 to 2018

```
      fillOpacity = 0.7,
      bringToFront = TRUE)) %>%
  addLegend(pal = pal,
            values  = map_data$cluster,
            na.label = "Missing data",
            position = "bottomleft",
            title = "Youth unemployment trajectories by NUTS 2 in Europe")
```

[24]:
```
Output in Figure 6
```

[25]:
```
# Calculate country summary statistics
freq_reg <- map_data@data %>%
  group_by(cntr_name, cluster) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

[26]:
```
# reformat the data to order them by clusters frequency
data_wide <- dcast(freq_reg, cntr_name ~ cluster, value.var="freq")
data_wide <- data_wide[order(-data_wide$`Stable Low youth unemployment`,
                             -data_wide$`Stable Moderate youth unemployment`,
                             -data_wide$`Increasingly High youth unemployment`,
                             -data_wide$`Stable High youth unemployment`),]
# Create a bar plot for country distribution of clusters
distribution_plot <- ggplot()
  geom_bar(aes(y=freq, x=cntr_name, fill=cluster), data=freq_reg, stat="identity")
  labs(title = "Distribution of youth unemployment trajectories across Europe",
       x = "Countries", y = "Proportion", fill = "")
  theme_minimal()
  theme(axis.text.x=element_text(angle = 90, hjust = 1))
  scale_x_discrete(limits=c(data_wide$cntr_name))
  scale_fill_brewer(palette="RdYlGn", na.value = "grey64", direction = -1)

# Set an interactive mode to the plot
ggplotly(distribution_plot)
```

[26]:
```
Output in Figure 7
```

6 Conclusion

Sequence analysis offers the opportunity to understand long-term socioeconomic trends over various levels of geographic regions. Clustering regions that follow similar socioeconomic trajectories can guide local, regional, national or European policy making by identifying and reducing the socioeconomic segregation of disadvantaged population groups. The first aim of this notebook was to highlight (NUTS 2) regions in Europe that maintain high, moderate or low youth unemployment levels, as well as regions that have transitioned from one level to another over the last decade. The findings of this notebook showed that northern Europe has high concentrations of regions with stable low youth unemployment, while southern Europe has high concentrations of regions with stable high youth unemployment. It is observed that southern countries struggled to adapt to the financial crisis of 2008. These findings can be used as a starting point to understand migration patterns that originated from these "disadvantaged" regions within or outside European boundaries. The second aim of this notebook was to provide a self-contained reproducible and transparent analytical workflow. This aim was achieved by providing detailed steps for successful data manipulation and make use of sequence analysis which is not a commonly used method in regional studies. Hence, I hope that data and regional scientists can benefit from the functionalities offered in the notebook and use it as a complementary guide when analysing their own data.

Acknowledgment

I would like to acknowledge the useful and constructive feedback received from my PhD supervisor Dr. Francisco Rowe throughout this research project. I would also like to thank my fellow PhD students at the University of Liverpool Krasen Samardzhiev and Patrick Ballantyne as well as the two anonymous reviewers for their useful comments on earlier versions of the notebook.

References

Bell DNF, Blanchflower DG (2011) Young people and the great recession. *Oxford Review of Economic Policy* 27[2]: 241–267. CrossRef.

Brzinsky-Fay C (2007) Lost in transition? Labour market entry sequences of school leavers in Europe. *European Sociological Review* 23[4]: 409–422. CrossRef.

Delmelle EC (2016) Mapping the DNA of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change. *Annals of the American Association of Geographers* 106[1]: 36–56. CrossRef.

Dietrich H (2012) Youth unemployment in Europe – Theoretical considerations and empirical findings. Friedrich ebert stiftung, bonn

Kaufman L, Rousseuw PJ (1991) Finding groups in data: An introduction to cluster analysis. Vol. 47. 2. CrossRef.

O'Reilly J, Eichhorst W, Gábos A, Hadjivassiliou K, Lain D, Leschke J, McGuinness S, Kureková LM, Nazio T, Ortlieb R, Russell H, Villa P (2015) Five characteristics of youth unemployment in Europe: Flexibility, education, migration, family legacies, and EU policy. *SAGE Open* 5[1]: 1–19. CrossRef.

Patias N, Rowe F, Cavazzi S (2020) A scalable analytical framework for spatio-temporal analysis of neighborhood change: A sequence analysis approach. In: Kyriakidis P, Hadjimitsis D, Skarlatos D, Mansourian A (eds), *Geospatial Technologies for Local and Regional Development*. Springer International Publishing, Cham, 223–241. CrossRef.

Peng RD (2011) Reproducible research in computational science. *Science* 334[6060]: 1226–1227. CrossRef.

Pop A, Kotzamanis B, Muller E, McGrath J, Walsh K, Peters M, Girejko R, Dietrich C (2019) YUTRENDS – Youth unemployment: Territorial trends and regional resilience. ESPON, Luxemburg

Rowe F, Casado-Díaz JM, Martínez-Bernabéu L (2017a) Functional labour market areas for Chile. *REGION* 4[3]: R7–R9. CrossRef.

Rowe F, Corcoran J, Bell M (2017b) The returns to migration and human capital accumulation pathways: Non-metropolitan youth in the school-to-work transition. *Annals of Regional Science* 59[3]: 819–845. CrossRef.

Rule A, Birmingham A, Zuniga C, Altintas I, Huang C, Knight R, Moshiri N, Nguyen MH (2019) Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Computational Biology* 15[7]. CrossRef.

Salmela-Aro K, Kiuru N, Nurmi J, Eerola M (2011) Mapping pathways to adulthood among finnish university students: Sequences, patterns, variations in family- and work-related roles. *Advances in Life Course Research* 16[1]: 25–41. CrossRef.

Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten simple rules for reproducible computational research. *PLoS Computational Biology* 9[10]. CrossRef.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 74[12]: 5463–5467. CrossRef.

Studer M, Ritschard G (2016) What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 179[2]: 481–511. CrossRef.

Resources

REGION
The Journal of ERSA
Powered by WU

Volume 6, Number 3, 2019, R1–R57
DOI: 10.18335/region.v6i3.267

journal homepage: region.ersa.org

REAT: A Regional Economic Analysis Toolbox for R

Thomas Wieland[1]

[1] Karlsruhe Institute of Technology, Karlsruhe, Germany

Received: 7 June 2019/Accepted: 4 November 2019

Abstract. Methods of regional economic analysis are widely used in regional and urban economics as well as in economic geography. This paper introduces the REAT (Regional Economic Analysis Toolbox) package for the programming environment R, which provides a collection of mathematical regional analysis methods in a user-friendly way. The focus is on the identification of regional inequality, beta and sigma convergence, measurement of agglomerations, point-based measures of clustering and accessibility, as well as regional growth. The theoretical basics of the applications are briefly introduced, while the usage of the most important functions is presented and explained using real data.

1 Introduction

Methods of regional economic analysis (or regional analysis) are used frequently in theory-based, empirical studies from regional and urban economics as well as (quantitative) economic geography. These methods aim at analyzing some of the most important issues in the mentioned research fields, including (but not limited to) the existence and evolution of agglomerations, regional economic growth and regional disparities (Capello, Nijkamp, 2009; Dinc, 2015; Farhauer, Kröll, 2014; Schätzl, 2000). In any of the mentioned fields, a growing amount of quantitative data has to be processed when using traditional or novel methods and models of regional analysis. This paper introduces the package (add-on) `REAT` (Regional Economic Analysis Toolbox) (Wieland, 2019) for the programming environment `R` (R Core Team, 2018a). The package provides a collection of mathematical regional analysis applications, designed in a relatively user-friendly way.

The main topics in the regional analysis context can be summarized as follows, showing also the structure of the present paper with respect to the presented approaches and their application in `REAT`:

1. Identifying regional inequality (or regional disparities) using indicators of concentration and/or dispersion (Section 2)
2. Regional disparities over time leading to the concept of beta and sigma convergence (Section 3)
3. Measuring agglomerations, which means the specialization of regions and the spatial concentration of industries as well as more complex cluster indices (Section 4)
4. Point-based measures of clustering and accessibility (Section 5)
5. Regional growth, especially shift-share analysis (Section 6)

Note that, in its original form, the open source software `R` is a command-line environment including a lot of mathematical and statistical features. For the installation of `R` and its packages as well as the basics of navigation and implemented statistical functions, see the `R` documentations (R Core Team, 2018b). A good supplement for working with `R` is `RStudio` (RStudio Team, 2016). The `REAT` package deals with several `R` data types: The most functions require and calculate `numeric vectors`, but, in some cases, also objects of type `matrix`, `data frame` and `list`, depending on the complexity of calculation. For a quick introduction to the data types in `R` and their properties, see e.g. Kabacoff (2017).

2 Concentration, dispersion and regional disparities

2.1 Indicators of concentration and dispersion

Regional disparities are a frequent topic in economic geography and regional economics. The spatial inequality with respect to e.g. regional output, income or employment is an essential element of polarization theory (Myrdal, 1957) and "New Economic Geography" (Krugman, 1991; Fujita et al., 2001). Assessing regional disparities is possible using concentration and dispersion indicators, which belong to the univariate and descriptive analysis in statistics. Apart from regional economics, these measures are used in several contexts, such as competition economics (market concentration of firms) or welfare economics (income inequality). For a review of the most common indicators with respect to regional inequality, see Portnov, Felsenstein (2010), for studies comparing different indicators in the regional economic context using empirical data, see e.g. Gluschenko (2018); Habánik et al. (2013); Huang, Leung (2009); Palan (2017); Petrakos, Psycharis (2016).

Concentration is operationalized as the discrepancy between an empirical distribution of a variable x (e.g. annual turnover, income, gross domestic product [GDP]) with n observations or objects (e.g. competing firms, households, regions) and a (theoretical) equal distribution or a reference distribution (e.g. population distribution). Dispersion indicators aim at the deviation from the arithmetic mean of x, $\bar{x}$. In this context, Portnov, Felsenstein (2005, 2010) distinguish between measures of deprivation and variation.

Typical measures of regional disparities are the Gini coefficient, the Herfindahl-Hirschman index and the coefficient of variation (Lessmann, 2005). The most popular measure of concentration is the Gini coefficient (Gini, 1912) in combination with the Lorenz curve (Lorenz, 1905). There are several calculation approaches for the Gini coefficient, all producing the same result. The Lorenz curve is a graphical indicator, showing the deviation of the empirical shares of the regarded variable x from a (theoretical) equal distribution. Another well-known indicator is the Herfindahl-Hirschman index, which was developed independently by Hirschman (1945) and Herfindahl (1950), both in the context of competition economics. Several other concentration indicators are also applied in the fields of regional economics with respect to regional disparities, such as the Hoover coefficient (Hoover, 1936) and the Theil coefficient (Theil, 1967).

Except for the standard deviation, whose unit is equal to the unit of x, all common indicators are dimensionless. Most of them (except for standard deviation and coefficient of variation) have a fixed value range, normally between zero (indicating complete equality/dispersion) and one (indicating complete inequality/concentration).

Most of the common indicators are mathematically formulated in an unweighted and in a weighted form, while, in the context of regional disparities, the latter is mostly done using the regions' proportion of the total (e.g. national) population (Doran, Jordan 2013; Lessmann 2014; Mussini 2017; Petrakos, Psycharis 2016; for a critical discussion of weighting these coefficients, see Gluschenko 2018). In the literature, there are different formulations where the weighted coefficients also include a weighted arithmetic mean. Note that, in the case of the population-weighted Gini coefficient, a weighted arithmetic mean is mandatory to keep the indicators' value range.

Especially when dealing with GDP per capita as an indicator of regional economic output, several recent studies use dispersion measures rather than concentration measures, especially the (weighted) coefficient of variation (e.g. Lessmann 2005, 2014, 2016;

Lessmann, Seidel 2017; Petrakos, Psycharis 2016). This dispersion indicator is a dimensionless normalization of the standard deviation. Weighting the coefficient of variation with population shares was introduced by Williamson (1965), which has led to calling this coefficient the Williamson index. As regional incomes or outputs are not normally distributed in most cases, resulting in biased arithmetic means used in the calculation of dispersion measures, the regarded variable may be log-transformed, which means replacing x_i with $\log(x_i)$ in the calculations.

Table 1 shows the common indicators, including their (population-)weighted and their normalized form (if there exist any) and the corresponding value ranges. The formulae are shown in a way that includes several ways of application. The regarded variable is always named x_i, while the (population) weighting is called w_i. Some indicators, such as the Hoover or the Coulter coefficient, require a variable representing a reference distribution the shares of x_i are compared to. This reference is *not* a weighting. However, in many studies, the regional population is also used for the reference distribution. In these cases, reference and weighting are the same data. The reference distribution may also be equal to $1/n$.

Several indicators are also used for the analysis of regional specialization or the spatial concentration of industries, such as the Hoover coefficient or the Herfindahl-Hirschman index or its inverse ($1/HHI$; also known as the "equivalent number" in the competition context). Other coefficients of concentration and specialization are discussed in Section 4. The last coefficient in Table 1, the mean square successive difference (von Neumann et al., 1941) is a measure for time variability not originating from but also transferable to regional economics.

2.2 Application in REAT

2.2.1 REAT functions for concentration and dispersion indicators

Table 2 shows the functions for concentration and dispersion measures implemented in the `REAT` package. All functions require at least one argument, a `numeric vector` with a length equal to n, containing the regarded variable x (e.g. income) with i observations (e.g. regions), where $i = 1, ..., n$. This data may be a single `vector` or a column of a `data frame` or `matrix`.

An optional weighting of the vector `x` can be done using the function argument `weighting` which is also a `numeric vector` of length n. By default, the functions remove missing (`NA`) values. The `hoover()` function always needs a reference distribution (see the Hoover coefficient formula in Table 1), which is stated via the `ref` argument, also requiring a `numeric vector` of length n. If no reference variable is stated (`ref = NULL`), the reference is set to $1/n$.

All functions (except for `disp()`) return the single value of the computed coefficient. In the relevant cases (`gini()`, `gini2()`, `herf()` and `cv()`), a normalization of the coefficient is possible using the function argument `coefnorm = TRUE`, returning the normalized coefficient instead of the raw coefficient. The function `disp()` is a wrapper for all mentioned functions, calculating all coefficients (except for the *MSSD*) at once for one vector `x` or a set of variables/columns from a `data frame` or `matrix`.

Note that there are two functions for the Gini coefficient, `gini()` and `gini2()`, both producing the same result in the unweighted case. The former function is designed for income inequality, where the `weighting` option is designed for the calculation of the Gini coefficient for groups (e.g. income classes), where the weighting represents the group mean. The function `gini2()` is designed for the population-weighted analysis of regional inequality.

2.2.2 Application example: Small-scale regional disparities in health care provision

Regional inequality with respect to health care providers is a topic of high societal significance. In Germany, the health care planning system (*Kassenärztliche Bedarfsplanung*) attempts to flatten the disparities of local health care provision (Kassenärztliche Bundesvereinigung, 2013). Here, we analyze small-scale regional disparities in health care

Table 1: Indicators of concentration and dispersion for analyzing regional disparities

Indicator	Unweighted	Weighted	Normalized
Gini	$G = \frac{1}{2n^2\bar{x}} \sum_{i=1}^{n} \sum_{j=1}^{n} \lvert x_i - x_j \rvert$ $0 \leq G \leq 1 - \frac{1}{n}$	$G^w = \frac{1}{2\bar{x}^w} \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \lvert x_i - x_j \rvert$ $0 \leq G \leq 1 - \frac{1}{n}$	$G^* = \frac{n}{n-1} G$ $0 \leq G^* \leq 1$
HHI	$HHI = \sum_{i=1}^{n} (\frac{x_i}{\sum_{i=1}^{n} x_i})^2$ $\frac{1}{n} \leq HHI \leq 1$		$HHI^* = \frac{HHI - \frac{1}{n}}{1 - \frac{1}{n}}$ $0 \leq HHI^* \leq 1$
Hoover	$HC = \frac{1}{2} [\sum_{i=1}^{n} \lvert \frac{x_i}{\sum_{i=1}^{n} x_i} - \frac{r_i}{\sum_{i=1}^{n} r_i} \rvert]$ $0 \leq HC \leq 1$	$HC^w = \frac{1}{2} [\sum_{i=1}^{n} w_i \lvert \frac{x_i}{\sum_{i=1}^{n} x_i} - \frac{r_i}{\sum_{i=1}^{n} r_i} \rvert]$ $0 \leq HC \leq 1$	
Theil	$TC = \frac{1}{n} \sum_{i=1}^{n} \ln(\frac{\bar{x}}{x_i})$ $0 \leq TC \leq 1$	$TC^w = \frac{1}{n} \sum_{i=1}^{n} w_i \ln(\frac{\bar{x}}{x_i})$ $0 \leq TC^w \leq 1$	
Coulter		$CC = \sqrt{\frac{1}{2} [\sum_{i=1}^{n} w_i (\frac{x_i}{\sum_{i=1}^{n} x_i} - \frac{r_i}{\sum_{i=1}^{n} r_i})^2]}$ $0 \leq CC \leq 1$	
Atkinson	$AI = 1 - [\frac{1}{n} \sum_{i=1}^{n} x_i^{1-\epsilon}]^{\frac{1}{1-\epsilon}}$ $0 \leq AI \leq 1$		
Dalton	$\delta = \frac{\log(\frac{1}{n} \sum_{i=1}^{n} x_i)}{\log(\sqrt[n]{\pi_{i=1}^{n} x_i})}$ $0 \leq \delta \leq \infty$		
SD	$s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ $0 \leq s \leq \infty$	$s^w = \sqrt{\frac{1}{n} \sum_{i=1}^{n} w_i (x_i - \bar{x})^2}$ $0 \leq s \leq \infty$	see CV
CV	$v = \frac{1}{\lvert \bar{x} \rvert} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ $0 \leq v \leq \infty$	see Williamson	$v^* = \frac{v}{\sqrt{n}}$ $0 \leq v^* \leq 1$
Williamson		$WI = \frac{1}{\lvert \bar{x} \rvert} \sqrt{\frac{1}{n} \sum_{i=1}^{n} w_i (x_i - \bar{x})^2}$ $0 \leq v \leq \infty$	
MSSD	$MSSD = \frac{\sum_{t=1}^{T-1} (x_{t+1} - x_t)^2}{T-1}$		

Notes: x_i is the i-th observation of the regarded variable x (e.g. GDP [per capita] in region i), x_j is the value of the same variable with respect to object j, r_i is the i-th observation of a reference variable (e.g. population), n is the number of objects (e.g. regions), $\bar{x}$ is the arithmetic mean of x: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{x}^w$ is the weighted arithmetic mean of x: $\bar{x}^w = \frac{1}{n} \sum_{i=1}^{n} w_i x_i$, w_i and w_j are the population weightings: $P_i / \sum_{i=1}^{n} P_i$ and $P_j / \sum_{j=1}^{n} P_j$, where P_i and P_j are the population sizes of regions i and j, respectively, ϵ is an inequality aversion parameter ($0 < \epsilon < \infty$) for the Atkinson index, t is a given time period and T is the number all regarded time periods.
Compiled from: Charles-Coll (2011); Cracau, Durán Lima (2016); Damgaard, Weiner (2000); Gluschenko (2018); Heinemann (2008); Kohn, Öztürk (2013); Portnov, Felsenstein (2005, 2010); Taylor, Cihon (2004); Schätzl (2000); Störmann (2009)

provision in two neighboring German counties (Göttingen and Northeim) using the data on medical practices and local population from Wieland, Dittrich (2016). The data is stored in the datasets `GoettingenHealth1` and `GoettingenHealth2`, both included as example datasets in the `REAT` package. The study area is segmented into 420 districts, representing either city districts of larger cities or villages and hamlets.

The dataset `GoettingenHealth2` contains these 420 regions with an individual ID

Table 2: REAT functions for concentration and dispersion indicators

Indicator	REAT function	Mandatory arguments	Optional arguments	Output
Gini/ Lorenz	gini()	vector x	weighting vector, remove NAs, Lorenz curve, normalization	value: G or G^* or G^w, optional: plot (LC)
	gini2()	vector x	weighting vector P_i, remove NAs, normalization	value: G or G^* or G^w,
	lorenz()	vector x	weighting vector, remove NAs,	plot LC, value: G or G^w and/or G^*
HHI	herf()	vector x	remove NAs, normalization	value: HHI or HHI^* or N_{HHI}
Hoover	hoover()	vector x reference vector r_i	weighting vector P_i, remove NAs	value: HC or HC^w
Theil	theil()	vector x	weighting vector P_i, remove NAs	value: TC or TC^w
Coulter	coulter()	vector x	weighting vector P_i, remove NAs	value: CC
Atkinson	atkinson()	vector x	remove NAs, epsilon	value: AI
Dalton	dalton()	vector x	remove NAs	value: δ
SD	sd2()	vector x	weighting vector, remove NAs, treating as sample	value: s or s^W
CV	cv()	vector x	weighting vector, remove NAs, normalization, treating as sample	value: v or v^W or v^*
Williamson	williamson()	vector x, weighting vector P_i	remove NAs	value: WI
MSSD	mssd()	vector x	remove NAs	value: $MSSD$
All indicators	disp()	vector x or vectors $x_1, x_2, ...$ from dataframe	weighting vector P_i, remove NAs	matrix with 13 (no weighting) or 19 indicators (incl. weighted)

Source: own compilation.

(column `district`) and geographic coordinates (columns `lat` and `lon`, respectively) and the number of general practitioners, psychotherapists and pharmacies located there (columns `phys_gen`, `psych` and `pharm`, respectively) as well as the local population (column `pop`). First, we load the dataset:

```
data(GoettingenHealth2)
```

Now, we investigate how the health care providers are dispersed over the whole area. In the first step, we calculate the Gini coefficient for the concentration of general practitioners using the `REAT` function `gini()`:

```
gini (GoettingenHealth2$phys_gen)
[1] 0.8386269
```

The empirical Gini coefficient is equal to 0.839, indicating a relatively strong concentration. If we want to calculate the normalized (unbiased) indicator instead, we use the same function with the optional argument `coefnorm = TRUE`:

```
gini (GoettingenHealth2$phys_gen, coefnorm = TRUE)
[1] 0.8406284
```

In the same way, we calculate e.g. the Herfindahl-Hirschman index, non-normalized and normalized:

```
herf (GoettingenHealth2$phys_gen)
[1] 0.01528053

herf (GoettingenHealth2$phys_gen, coefnorm = TRUE)
[1] 0.01293036
```

Remember that the minimum of HHI is $1/n$ (here: $1/420 \approx 0.00238$) and the minimum of HHI^* is equal to zero.

If we want to inspect the concentration graphically, we could use the Lorenz curve, which can be plotted using either the functions `gini()` or `lorenz()`. Here, we use `gini()`, tell the function to plot the curve (`lc = TRUE`), and include several graphical parameters (such as `lc.col` for the color of the Lorenz curve or `lcx` and `lcy` for the x/y axes labels). As we want to compare the population distribution to the location distribution, we start by plotting the Lorenz curve for the local population:

```
gini(GoettingenHealth2$pop, lc = TRUE, lsize = 1, le.col = "black",
lc.col = "orange", lcx = "Shares of districts", lcy = "Shares of
providers", lctitle = "Spatial concentration of health care
providers", lcg = TRUE, lcgn = TRUE, lcg.caption =
"Population 2016:", lcg.lab.x = 0, lcg.lab.y = 1)
# Gini coefficient and Lorenz curve for the no. of inhabitants
[1] 0.5840336
```

Now, we overlay the Lorenz curves of general practitioners and psychotherapists, which means adding two more curves (function argument `add.lc = TRUE`):

```
gini(GoettingenHealth2$phys_gen, lc = TRUE, lsize = 1, add.lc = TRUE,
lc.col = "red", lcg = TRUE, lcgn = TRUE, lcg.caption =
"Physicians 2016:", lcg.lab.x = 0, lcg.lab.y = 0.85)
# Adding Gini coefficient and Lorenz curve for the general practitioners
[1] 0.8386269

gini(GoettingenHealth2$psych, lsize = 1, lc = TRUE, add.lc = TRUE,
lc.col = "blue", lcg = TRUE, lcgn = TRUE, lcg.caption =
"Psychotherapists 2016:", lcg.lab.x = 0, lcg.lab.y = 0.7)
# Adding Gini coefficient and Lorenz curve for psychotherapists
[1] 0.9329298
```

Our commands result in the output of Figure 1, showing three Lorenz curves (population, general practitioners and psychotherapists) and the line of equality (diagonal). All three empirical distributions differ from an equal distribution. In about 72% of the regions, representing about 23% of the whole population (orange curve; $G \approx 0.584$), no general practitioner is located (red curve; $G \approx 0.839$). But the psychotherapists are more concentrated, as they are located only in about 13% of all districts (blue curve; $G \approx 0.933$). As we can see, the physicians are more concentrated than the inhabitants but the psychotherapists are more concentrated than the physicians.

Now, we calculate all mentionened concentration and dispersion coefficients at once for all three types of providers using the function `disp()`, including a population weighting:

```
disp(GoettingenHealth2[c(5,6,7)], weighting = GoettingenHealth2$pop)
# column 5 = general practitioners, column 6 = psychotherapists,
# column 7 = pharmacies, column "pop" = local population
```

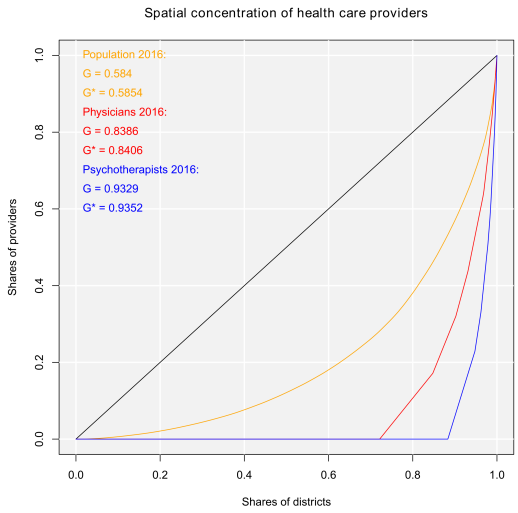


Figure 1: Lorenz curves for the spatial concentration of health care providers

Our output is:

```
Concentration and dispersion measures
Note: w = weighted, n = normalized, eq = equivalent number

                phys_gen         psych         pharm
Gini        0.838626907  0.932929782  0.891547619
Gini n      0.840628403  0.935156345  0.893675418
Gini w      0.629454516  0.770895945  0.705628058
Gini w n    0.630956794  0.772735792  0.707312135
HHI         0.015280527  0.038494685  0.024166667
HHI n       0.012930361  0.036199923  0.021837709
HHI eq     65.442769020 25.977611940 41.379310345
Hoover      0.721428571  0.883333333  0.838095238
Hoover w    0.001852337  0.003130602  0.003418787
Theil                NA           NA           NA
Theil w              NA           NA           NA
Coulter     0.049850824  0.123305927  0.065569205
Atkinson    0.761164110  0.900755425  0.854223763
Dalton               NA           NA           NA
SD          1.714506606  1.095496987  0.865286915
SD w        4.010246439  1.847716870  2.401476794
CV          2.330397328  3.899226565  3.028504203
CV n        0.113847359  0.190489683  0.147952112
Williamson  1.429449565  1.965446423  1.709288672
```

We conclude that any concentration/dispersion measure is the highest for psychotherapists and the lowest for the general practitioners, while the values for pharmacies lie between them. The regional disparities with respect to pharmacies are higher than those with respect to general practitioners, while the most unequal distribution is that of psychotherapists. In other words: The pharmacies are more spatially concentrated than the general practitioners and the psychotherapists are the most concentrated health locations here.

In most cases, population weighting reduces the coefficient values. That is, because districts with a large (small) population have a high (low) impact on the resulting coefficient and the districts without health service providers are also small districts. Furthermore, as the regarded variables contain zero values (which means no health service locations), the Theil coefficient (including the term $\ln(\bar{x}/x_i)$) and the Dalton coefficient (including the n-th root) cannot be computed, resulting in an output of `NA`.

The visible output of any function presented above can be saved in a new `R` object:

```
gini_phys <- gini (GoettingenHealth2$phys_gen)
# save as gini_phys (numeric vector of length = 1)
```

We can simply access our result:

```
gini_phys
[1] 0.8386269
```

The function `disp()` returns a `matrix` with 13 rows (when only unweighted coefficients are computed) or 19 rows (in the case of additional weighted coefficients) and one column for each regarded variable:

```
disp_Goettingen <- disp(GoettingenHealth2[c(5,6,7)],
weighting = GoettingenHealth2$pop)
# save as disp_Goettingen (matrix)
```

We call our results:

```
disp_Goettingen

                phys_gen       psych       pharm
Gini          0.83862691  0.93292978  0.89154762
Gini n        0.84062840  0.93515634  0.89367542
...
```

3 Regional convergence

3.1 The concept of beta and sigma convergence

Regional convergence is derived from (regional) growth theory (for an extensive survey, see Barro, Sala-i Martin 2004) and means the decline of regional disparities *over time*. The neoclassical growth model states that a region's economic output (e.g. GDP per capita) depends on its stock of factors of production, capital and labor (aggregate production function), on condition of constant returns to scale and diminishing marginal product of the factor inputs. As a consequence, regions with a high (low) initial level of factor input grow slower (faster) than "poor" ("rich") regions, what is called beta convergence. It is assumed that all regions converge to the same regional output level (steady-state). Sigma convergence means the decline of regional inequality with respect to regional output over time itself (Allington, McCombie, 2007; Capello, Nijkamp, 2009).

Both types of convergence can be tested empirically, as presented in Table 3. When testing for beta convergence, the natural logarithms of output growth over T time periods in i regions is regressed against the natural logarithms of the initial output values at time t. The original convergence formula was presented by Barro, Sala-i Martin (2004) using a nonlinear least squares (NLS) estimation approach. But in many cases, a linear transformation is used which allows for ordinary least squares (OLS) estimation (Allington, McCombie, 2007; Dapena et al., 2016; Schmidt, 1997; Young et al., 2008). The outcome variable of the convergence equation can be the regional growth between two years (e.g. Young et al. 2008) or the average growth rate per year (e.g. Goecke, Hüther 2016; Puente 2017; Weddige-Haaf, Kool 2017). Significance tests are carried out with t-tests for the regression coefficients and, in the OLS case, the F-test for the significance of R^2.

The estimated parameter of interest is the slope of the model, here denoted β (that is why the modeled process is called *beta* convergence): If $\beta < 0$ and statistically significant, there is *absolute* beta convergence. If additional variables (conditional variables) are included into the convergence equation, we have a test for *conditional* beta convergence. A further interpretation of the β coefficient is possible using the speed of convergence, λ, and H, the so-called half-life, which means the time (measured in the regarded time periods) to reduce the regional disparities by one half (Allington, McCombie, 2007; Schmidt, 1997).

Sigma convergence (which is named after the Greek letter for the standard deviation, σ) can be tested in two ways depending on the number of time periods: The regional

Table 3: Beta and sigma convergence

Type of convergence	Two time periods	More than two time periods
Beta convergence	absolute	
and estimation type	NLS	NLS
	$\frac{1}{T}\ln(\frac{Y_{i,t2}}{Y_{i,t1}}) = \alpha - [\frac{(1-e^{-\beta T})}{T}]\ln(Y_{i,t1}) + \epsilon$	$\frac{1}{T}\sum_{t=1}^{T}\ln(\frac{Y_{i,t+1}}{Y_{i,t}}) = \alpha - [\frac{(1-e^{-\beta T})}{T}]\ln(Y_{i,t1}) + \epsilon$
	OLS	OLS
	$\frac{1}{T}\ln(\frac{Y_{i,t2}}{Y_{i,t1}}) = \alpha + \beta\ln(Y_{i,t1}) + \epsilon$	$\frac{1}{T}\sum_{t=1}^{T}\ln(\frac{Y_{i,t+1}}{Y_{i,t}}) = \alpha + \beta\ln(Y_{i,t1}) + \epsilon$
	conditional	
	NLS	NLS
	$\frac{1}{T}\ln(\frac{Y_{i,t2}}{Y_{i,t1}}) = \alpha - [\frac{(1-e^{-\beta T})}{T}]\ln(Y_{i,t1}) + \theta X_i + \epsilon$	$\frac{1}{T}\sum_{t=1}^{T}\ln(\frac{Y_{i,t+1}}{Y_{i,t}}) = \alpha - [\frac{(1-e^{-\beta T})}{T}]\ln(Y_{i,t1}) + \theta X_i + \epsilon$
	OLS	OLS
	$\frac{1}{T}\ln(\frac{Y_{i,t2}}{Y_{i,t1}}) = \alpha + \beta\ln(Y_{i,t1}) + \theta X_i + \epsilon$	$\frac{1}{T}\sum_{t=1}^{T}\ln(\frac{Y_{i,t+1}}{Y_{i,t}}) = \alpha + \beta\ln(Y_{i,t1}) + \theta X_i + \epsilon$
	$\beta < 0$	$\beta < 0$
	Convergence speed: $\lambda = \frac{-\ln(1+\beta)}{T}$	
	Half-life: $H = \frac{\ln(2)}{\lambda}$	
Sigma convergence	$\sigma_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{i,t} - \bar{Y}_t)^2}$ or	
	$cv_t = \frac{\sigma_t}{\lvert\bar{Y}_t\rvert}$	
	$\frac{\sigma_{t1}}{\sigma_{t2}} > 1$ or	$\sigma = a + bt + \epsilon$ or
	$\frac{cv_{t1}}{cv_{t2}} > 1$	$cv = a + bt + \epsilon$
	Test statistic: $\frac{\sigma^2_{t1}}{\sigma^2_{t2}}$	$b < 0$

Notes: $Y_{i,t}$ is the regional output (e.g. GDP per capita) of region i at time t, $\bar{Y}_t$ is the arithmetic mean of $Y_{i,t}$ for all regions at time t, T is the number of regarded time periods (e.g. years), X_i is a set of other variables (conditions), σ_t is the standard deviation of the regional output of all regions, cv_t is the corresponding coefficient of variation, α, β, θ, a and b are estimated coefficients, ϵ is an error term and n is the number of regions.
Compiled from: Allington, McCombie (2007); Barro, Sala-i Martin (2004); Furceri (2005); Schmidt (1997)

inequality between all regions at time t is measured using the standard deviation, σ_t, or the coefficient of variation, cv_t, for the GDP per capita in its original or natural-logged form. If only two years are regarded, the quotient of both parameters is computed. If e.g. $\sigma_{t1} > \sigma_{t2}$, the regional inequality has declined from $t1$ to $t2$. A significance test can be applied with a simple ANOVA (analysis of variance), where the test statistic is the quotient of the underlying variances (σ^2) (Furceri, 2005; Schmidt, 1997; Young et al., 2008). Within a time series, the dispersion parameter is regressed (and plotted) against time. If the slope coefficient of time is negative, there is sigma convergence (Goecke, Hüther, 2016; Huang, Leung, 2009; Schmidt, 1997).

3.2 Application in REAT

3.2.1 REAT functions for beta and sigma convergence

Table 4 shows the functions for beta and sigma convergence as implemented in `REAT`. The analysis of beta convergence is provided by the functions `betaconv.ols()` and `betaconv.nls()` for OLS and NLS estimation, respectively. Speed of convergence and

Table 4: REAT functions for beta and sigma convergence

Convergence	REAT function	Mandatory arguments	Optional arguments	Output
Beta convergence	betaconv.ols()	vectors $Y_{i,t1}$ and $Y_{i,t2}, ..., Y_{i,T}$, t_1 and t_T	Conditions, scatterplot	visible: model estimates, invisible: list with model estimates and regression data, optional: plot
	betaconv.nls()	vectors $Y_{i,t1}$ and $Y_{i,t2}, ..., Y_{i,T}$, t_1 and t_T	Conditions, scatterplot	visible: model estimates, invisible: list with model estimates and regression data, optional: plot
	betaconv.speed()	values β and T		matrix with λ and H
Sigma convergence	sigmaconv() (when $T = 2$)	vectors $Y_{i,t1}$ and $Y_{i,t2}$, t_1 and t_T	Sigma measure, log, weighting, normalization	visible: estimates, invisible: matrix with estimates
	sigmaconv.t() (when $T > 2$)	vectors $Y_{i,t1}$ and $Y_{i,t2}, ..., Y_{i,T}$, t_1 and t_T	Sigma measure, log, weighting, normalization, line plot	visible: model estimates, invisible: matrix with model estimates, optional: plot
All at once: Beta and sigma convergence	rca()	vectors $Y_{i,t1}$ and $Y_{i,t2}, ..., Y_{i,T}$, t_1 and t_T	Beta estimation, conditions, scatterplot, sigma measure, log, weighting, line plot	visible: model estimates, invisible: list with model estimates and regression data, optional: plot

Source: own compilation.

half-life can be computed with the function `betaconv.speed()`. The ratio test of sigma convergence for two time periods can be done using the function `sigmaconv()`, while a trend regression over time is implemented into the function `sigmaconv.t()`. Both convergence types can be analyzed at once with the function `rca()`, which is a wrapper for all functions mentioned above.

The functions require (at least) two `numeric vectors`, containing the regarded variable Y (e.g. GDP per capita) for at least two different time periods, e.g. from the same `data frame`. Also the start and end time periods (t_1 and t_T) have to be stated. Optionally, a graphical output can be generated (scatterplot for beta convergence, line plot for sigma convergence with respect to longitudinal data). Furthermore, when analyzing sigma convergence, the user can choose whether Y should be log-transformed or not and/or which sigma measure is computed (variance, standard deviation or coefficient of variation; weighted or non-weighted).

Note that, unlike the functions for regional inequality indicators (Section 2), the `REAT` functions for regional convergence distinguish between a *visible* and an *invisible* output. The latter can be saved as a new `R` object. While the visible output shows the main results, the invisible output goes beyond that: `betaconv.ols()`, `betaconv.nls()` and `rca()` return a `list`, which is the most flexible data type in `R`, because it consists of a non-predetermined number of different data objects. Apart from the model results, e.g. the (transformed) regression data is returned in this invisible output.

3.2.2 Application example: Beta and sigma convergence in Germany on the county level

In this example, we look at regional convergence in Germany. The `REAT` package includes the example dataset `G.counties.gdp` with the GDP (gross domestic product), the population and the GDP per capita for the 402 counties ("Kreise") in Germany 1992 to 2014

(complete data only for 2000-2014). First, we load the dataset:

```
data (G.counties.gdp)
```

In our case, we prevent scientific notation of numbers in R and set a limit of 4 digits:

```
options(scipen = 100, digits = 4)
```

We need the columns named gdppcxxxx, containing the GDP per capita for each year, e.g. G.counties.gdp$gdppc2010 contains the GDP per capita for 2010. In the first step, we test absolute beta convergence comparing the years 2010 and 2014 with OLS estimation using the function betaconv.ols():.

```
betaconv.ols (G.counties.gdp$gdppc2010, 2010, G.counties.gdp$gdppc2014,
2014, output.results = TRUE)
# Two years, no conditions (Absolute beta convergence)
```

The output is:

```
Absolute Beta Convergence
Model coefficients (Estimation method: OLS)
           Estimate Std. Error t value     Pr (>|t|)
Alpha      0.104159   0.018934   5.501 0.00000006743
Beta      -0.007373   0.001848  -3.990 0.00007867475
Lambda     0.001850         NA      NA            NA
Halflife 374.640507         NA      NA            NA
Model summary
          Estimate F value df 1 df 2    Pr (>F)
R-Squared  0.03827   15.92    1  400 0.00007867
```

We see that both regression coefficients, α and β, are statistically significant ($t \approx 5.50$ and -3.99, respectively, both $p < 0.001$) and the linear regression model is significant as a whole ($F \approx 15.92$, $p < 0.001$). The negative sign of β shows that, on average, the higher the initial GDP per capita, the lower its growth, which indicates absolute beta convergence. However, the convergence process is very slow: The speed of convergence, represented by λ, shows a harmonization by 0.185% per year. This implies that the output gap will be reduced by 50% in approximately 375 years.

Now we check sigma convergence for the same time using the function sigmaconv(). We choose the coefficient of variation as measure, while using the GDP per capita values in their original form:

```
sigmaconv (G.counties.gdp$gdppc2010, 2010, G.counties.gdp$gdppc2014,
2014, sigma.measure = "cv", output.results = TRUE)
# Using the coefficient of variation
```

The output is:

```
Sigma convergence for two periods (ANOVA)
         Estimate F value df1 df2 Pr (>F)
CV 2010   0.03416      NA  NA  NA      NA
CV 2014   0.03316      NA  NA  NA      NA
Quotient  1.03004   1.038 401 401  0.7117
```

The coefficient of variation is a little smaller in 2014, which means the spatial inequality declined between 2010 and 2014. The quotient of the variances is slightly above one ($F = \sigma^2_{2010}/\sigma^2_{2014} \approx 1.04$), but not statistically significant ($p \approx 0.71$).

When analyzing regional convergence with REAT, it is preferable (and more convenient) to use the wrapper function rca(). Instead of repeating the results above, we test for (absolute) beta and sigma convergence between 2000 and 2014. The analysis of sigma convergence uses trend regression (function argument sigma.type = "trend") for the coefficient of variation (sigma.measure = "cv"). We also want plots for both convergence types (beta.plot = TRUE and sigma.plot = TRUE, respectively) with specific axis labels (e.g. beta.plotX = "Ln (initial GDP p.c.)"). Our code is:

```
rca (G.counties.gdp$gdppc2000, 2000, G.counties.gdp[55:68], 2014,
conditions = NULL, sigma.type = "trend", sigma.measure = "cv",
beta.plot = TRUE, beta.plotLine = TRUE, beta.plotX =
"Ln (initial GDP p.c.)", beta.plotY = "Ln (av. growth GDP p.c.)",
beta.plotTitle = "Beta convergence of German counties 2000-2014",
sigma.plot = TRUE, sigma.plotY = "cv of ln (GDP p.c.)",
sigma.plotTitle = "Sigma convergence of German counties 2000-2014")
# 14 years: 2000 (column 55) to 2014 (column 68)
# no conditions (Absolute beta convergence)
# with plots for both beta and sigma convergence
```

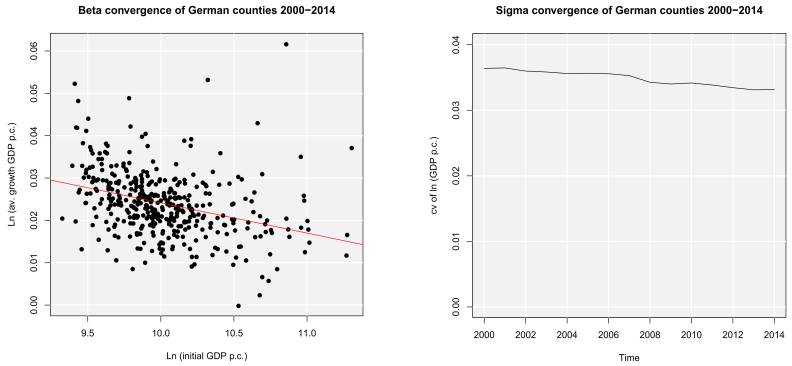
This results in the following output:

```
Regional Beta and Sigma Convergence

Absolute Beta Convergence
Model coefficients (Estimation method: OLS)
             Estimate Std. Error t value                  Pr (>|t|)
Alpha       0.0954564  0.0099087   9.634 0.000000000000000000006845
Beta       -0.0071323  0.0009885  -7.215 0.000000000002719258225550
Lambda      0.0005113         NA      NA                          NA
Halflife 1355.7282963         NA      NA                          NA
Model summary
          Estimate F value df 1 df 2           Pr (>F)
R-Squared   0.1152   52.06    1  400 0.000000000002719

Sigma convergence (Trend regression)
            Estimate Std. Error t value        Pr(>|t|)
Intercept  0.5523659 0.03084855   17.91 0.0000000001526
Time      -0.0002579 0.00001537  -16.78 0.0000000003446
Model summary
          Estimate F value df 1 df 2         Pr (>F)
R-Squared   0.9558   281.4    1   13 0.0000000003446
```

This function also produces the plots in Figures 2a and 2b, both showing a declining curve, which is a first indication of both beta and sigma convergence. The beta convergence model is statistically significant ($F \approx 52.06$, $p < 0.001$), as well as the coefficients α ($t \approx 9.63$, $p < 0.001$) and β ($t \approx -7.21$, $p < 0.001$). Again, we find evidence for absolute beta convergence because of a negative slope ($\beta \approx -0.007$). The trend regression model for sigma convergence is significant ($F \approx 281.4$, $p < 0.001$). The slope is significant and negative ($b \approx -0.00026$, $t \approx 17.91$, $p < 0.001$), which indicates sigma convergence. However, both types of convergence can be regarded as very slow processes: The half-life value shows that, resulting from the beta convergence model, the regional disparities in GDP per capita will be halved in approximately 1,356 years. When looking at the trend regression, we see that the coefficient of variation declines only by 0.00026 per year. Another aspect is that we only regarded absolute beta convergence, ignoring other spatial effects or the impact of regional policy. The latter is also not considered in neoclassical regional growth theory.

Remembering German reunification, we want to test if there are average growth differences between West Germany and East Germany (former German Democratic Republic), which leads to conditional beta convergence. The dataset `G.regions.emp` contains the column `regional`, where the counties are attributed either to West or East Germany, expressed as character string (`"West"` or `"East"`). We need to include our condition into the convergence equation. Thus, we use the `REAT` function `to.dummy()` to create dummy variables (1/0) out of (nominal scaled) variables, and add the indicator for West Germany (1, otherwise 0) to our data:

```
regionaldummies <- to.dummy(G.counties.gdp$regional)
# Creating dummy variables for West/East
# regionaldummies[,1] = East (1/0), regionaldummies[,2] = West (1/0)
G.counties.gdp$West <- regionaldummies[,2]
# Adding the dummy variable for West
```



(a) Absolute beta convergence (b) Sigma convergence

Figure 2: Regional convergence in Germany 2000-2014 (n = 402 counties)

Now, we test for conditional beta and sigma convergence, including the condition "West", again using the `rca()` function, but without plots and using the standard deviation (default setting) instead of the cv for sigma convergence. This time, we save the results in an object:

```
converg_results <- rca (G.counties.gdp$gdppc2000, 2000,
G.counties.gdp[55:68], 2014, conditions = G.counties.gdp[c(70)],
sigma.type = "trend")
# condition variable "West" in column 70
# Store results in "converg_results"
```

The output is:

```
Regional Beta and Sigma Convergence

Absolute Beta Convergence
Model coefficients (Estimation method: OLS)
             Estimate Std. Error t value                  Pr (>|t|)
Alpha       0.0954564  0.0099087   9.634 0.000000000000000000006845
Beta       -0.0071323  0.0009885  -7.215 0.000000000002719258225550
Lambda      0.0005113         NA      NA                          NA
Halflife 1355.7282963         NA      NA                          NA
Model summary
          Estimate F value df 1 df 2           Pr (>F)
R-Squared   0.1152   52.06    1  400 0.000000000002719

Conditional Beta Convergence
Model coefficients (Estimation method: OLS)
             Estimate Std. Error t value          Pr (>|t|)
Alpha       0.0754412  0.0102354   7.371 0.0000000000009872
Beta       -0.0047020  0.0010517  -4.471 0.0000101720129094
West       -0.0053559  0.0009745  -5.496 0.0000000693910790
Lambda      0.0003366         NA      NA                 NA
Halflife 2058.9555949         NA      NA                 NA
Model summary
          Estimate F value df 1 df 2                Pr (>F)
R-Squared   0.1774   43.04    2  399 0.00000000000000001192

Sigma convergence (Trend regression)
           Estimate Std. Error t value      Pr(>|t|)
Intercept  3.895236  0.3267817   11.92 0.00000002264
Time      -0.001764  0.0001628  -10.84 0.00000007041
Model summary
          Estimate F value df 1 df 2       Pr (>F)
R-Squared   0.9003   117.4    1   13 0.00000007041
```

In the `rca()` output, we can compare the results of absolute and conditional beta convergence. In the conditional model, the explained variance increases from $R^2 \approx 0.12$ to $R^2 \approx 0.18$, which indicates an increased explanatory power of the model due to the added condition variable. Both models are statistically significant, also the β values are negative and significant ($p < 0.001$ in both cases). The condition "West" is significant ($t \approx -5.50$, $p < 0.001$) and negative, which means that, on average, the GDP per capita in West German counties grew slower than in East Germany. These results *seem* to support the convergence hypothesis from growth theory, but one should not forget that e.g. political aspects (such as the German and/or EU regional policy) are not considered in this simple analysis.

As we have saved the invisible function output, we can access specific parts of our analysis, such as the regression data for the absolute convergence model:

```
converg_results$betaconv$regdata
# All results in list converg_results
# converg_results contains list betaconv (beta convergence results)
# betaconv contains data frame regdata (regression data)

    ln_initial   ln_growth
1       11.002  0.01997436
2       10.552  0.02980133
3       10.283  0.01794207
4       10.090  0.01763444
5       10.287  0.02361006
...
```

If we want to look at the single sigma values, we can address them via:

```
converg_results$sigmaconv$sigma.trend
# All results in list converg_results
# converg_results contains list sigmaconv (sigma convergence results)
# sigmaconv contains data frame sigma.trend (sigma values)

          years sigma.years
gdp1       2000      0.3646
gdppc2001  2001      0.3662
gdppc2002  2002      0.3618
gdppc2003  2003      0.3606
gdppc2004  2004      0.3592
...
```

4 Specialization of regions and spatial concentration of industries

4.1 Indicators of regional specialization and industry concentration

Specialization of regions or countries and the spatial concentration of industries or firms are phenomena linked to several research fields in regional economics and economic geography: Specialization is a key point in traditional theories of international trade with respect to comparative advantages (Ricardo, 1821) as well as in the generation of the "New Trade Theory" (introduced by Krugman 1979). Spatial clustering of firms or industries due to agglomeration economies is a perennial issue in all spatial economic fields. It especially reemerged in the context of the "New Economic Geography" (e.g. Krugman 1991; Fujita et al. 2001) as well as through the work of Porter (1990) regarding clusters. The common indicators are broadly discussed in Farhauer, Kröll (2014) or Nakamura, Morrison Paul (2009). For studies comparing some different indicators, see e.g. Goschin et al. (2009); Moga, Constantin (2011); Palan (2017).

When looking at the family of indicators of regional specialization and industry concentration, we have to distinguish between indicators for aggregate data, such as regional employment data, and those requiring individual firm data. The first group, compiled in Table 5, can be differentiated into indicators of specialization and indicators of spatial concentration. As both types of agglomeration are closely linked to each other, so are the

Table 5: Coefficients of regional specialization and industry concentration

Indicator	Specialization of region j	Spatial concentration of industry i
Hoover/Balassa	$LQ_{ij} = \frac{e_{ij}/e_i}{e_j/e} \equiv MRCA_{ij} = \frac{e_{ij}/e_j}{e_i/e}$	
	$\overline{LQ_j} = \frac{1}{I}\sum_{i=1}^{I} LQ_{ij}$	$\overline{LQ_i} = \frac{1}{J}\sum_{j=1}^{J} LQ_{ij}$
Extensions:		
O'Donoghue-Gleave	$SLQ_{ij} = \frac{LQ_{ij}-\overline{LQ_i}}{sd(LQ_i)}$	
Tian	$SLLQ_{ij} = \frac{\log(LQ_{ij})-\overline{\log(LQ_i)}}{sd(\log(LQ_i))}$	
Hoen-Oosterhaven	$ARCA_{ij} = \frac{e_{ij}}{e_j} - \frac{e_i}{e}$	
Hoover	$H_j = \frac{1}{2}[\sum_{i=1}^{I} \lvert\frac{e_{ij}}{e_j} - \frac{e_i}{e}\rvert]$ $0 \leq H_j \leq 1$	$H_i = \frac{1}{2}[\sum_{j=1}^{J} \lvert\frac{e_{ij}}{e_i} - \frac{e_j}{e}\rvert]$ $0 \leq H_i \leq 1$
Gini	$G_j = \frac{2}{I^2\bar{R}}\sum_{i=1}^{I} \lambda_i(R_i - \bar{R})$ $0 \leq G_j \leq 1$	$G_i = \frac{2}{J^2\bar{C}}\sum_{j=1}^{J} \lambda_j(C_j - \bar{C})$ $0 \leq G_i \leq 1$
	where: $R_i = \frac{e_{ij}/e_j}{e_i/e}$,	where: $C_j = \frac{e_{ij}/e_i}{e_j/e}$,
	$\bar{R} = \frac{1}{I}\sum_{i=1}^{I} R_i$ and	$\bar{C} = \frac{1}{J}\sum_{j=1}^{J} C_j$ and
	$\lambda_i = 1, ..., I$ $(\lambda_i < \lambda_{i+1})$	$\lambda_j = 1, ..., J$ $(\lambda_j < \lambda_{j+1})$
Krugman $(J = 2,\ I = 2)$	$K_{jl} = \sum_{i=1}^{I} \lvert s^s_{ij} - s^s_{il}\rvert$ $0 \leq K_{jl} \leq 2$	$K_{iu} = \sum_{j=1}^{J} \lvert s^c_{ij} - s^c_{uj}\rvert$ $0 \leq K_{iu} \leq 2$
	where: $s^s_{ij} = \frac{e_{ij}}{e_j}$ and $s^s_{il} = \frac{e_{il}}{e_l}$	where: $s^c_{ij} = \frac{e_{ij}}{e_i}$ and $s^c_{uj} = \frac{e_{uj}}{e_i}$
Extensions:		
Midelfart et al., Vogiatzoglou	$K_j = \sum_{i=1}^{I} \lvert s^s_{ij} - \bar{s}^s_{il}\rvert$ $0 \leq K_j \leq 2$	$K_i = \sum_{j=1}^{J} \lvert s^c_{ij} - \bar{s}^c_{uj}\rvert$ $0 \leq K_i \leq 2$
$(J > 2,\ I > 2)$	where: $s^s_{ij} = \frac{e_{ij}}{e_j}$ and	where: $s^c_{ij} = \frac{e_{ij}}{e_i}$ and
	$\bar{s}^s_{il} = \frac{1}{J-1}\sum_{i}^{J} s^s_{il}$, $l \neq j$	$\bar{s}^c_{uj} = \frac{1}{I-1}\sum_{u}^{I} s^c_{uj}$, $u \neq i$
Duranton-Puga	$RDI_j = \frac{1}{\sum_{i=1}^{I} \lvert\frac{s^s_{ij}}{s_i} - s_i\rvert}$ where: $s^s_{ij} = \frac{e_{ij}}{e_j}$ and $s_i = \frac{e_i}{e}$	
Litzenberger-Sternberg	$CI_{ij} = \frac{IS_{ij}ID_{ij}}{PS_{ij}}$ where $IS_{ij} = \frac{e_{ij}/a_j}{e_i/a}$, $ID_{ij} = \frac{e_{ij}/p_j}{e_i/p}$ and $PS_{ij} = \frac{e_{ij}/b_{ij}}{e_i/b_i}$	

Notes: e_{ij} and e_{il} equal the employment of industry i in regions j and l, respectively, e_i is the total employment in industry i, e_{uj} ist the employment of industry u in region j, e_j is the total employment in region j, e is the total employment in the whole economy, I is the number of industries, J is the number of regions, a_j is the area of region j, a is the total area in the whole economy, p_j is the population in region j, p is the total population, b_{ij} is the number of firms of industry i in region j and b_i is the number of firms in industry i.
Compiled from: Farhauer, Kröll (2014); Hoen, Oosterhaven (2006); Hoffmann et al. (2017); Nakamura, Morrison Paul (2009); O'Donoghue, Gleave (2004); Tian (2013); Schätzl (2000); Störmann (2009)

corresponding indicators. The empirical basis of all those measures is the employment e in industry i in region j, e_{ij}. This employment stock is compared to some reference, mostly including the total employment in region j, e_j, and/or the total employment in industry i, e_i, as well as the all-over employment e. The individual firm level indicators in Table 6 can be segmented into indicators for agglomeration of *one* industry due to localization economies and indicators for the coagglomeration of *different* industries due to urbanization economies.

Table 6: Coefficients of agglomeration and coagglomeration using individual firm data

Indicator	Agglomeration	Coagglomeration
Ellison-Glaeser	$\gamma_i = \frac{G_i - (1-\sum_{j=1}^{J} s_j^2) HHI_i}{(1-\sum_{j=1}^{J} s_j^2)(1-HHI_i)}$ where: $G_i = \sum_{j=1}^{J} (s_{ij}^c - s_j)^2$, $s_{ij}^c = \frac{e_{ij}}{e_i}$, $s_j = \frac{e_j}{e}$ and $HHI_i = \sum_{k=1}^{K} (\frac{e_{ik}}{e_i})^2$ z-standardization: $z_i = \frac{G_i - (1-\sum_{j=1}^{J} s_j^2) HHI_i}{\sqrt{var(G_i)}}$ where: $var(G_i) = 2\Big\{ HHI_i^2 \Big[\sum_{j=1}^{J} s_j^2$ $-2\sum_{j=1}^{J} s_j^3 + (\sum_{j=1}^{J} s_j^2)^2 \Big] -$ $\sum_{k=1}^{K} z_{ik}^4 \Big[\sum_{j=1}^{J} s_j^2 - 4\sum_{j=1}^{J} s_j^3$ $+3(\sum_{j=1}^{J} s_j^2)^2 \Big] \Big\}$	$\gamma^c = \frac{G/(1-\sum_{j=1}^{J} s_j^2) - HHI_U - \sum_{i=1}^{U} \gamma_i s_i^2 (1-HHI_i)}{1-\sum_{i=1}^{U} s_i^2}$ where: $G = \sum_{j=1}^{J} (x_j - s_j)^2$, $x_j = \sum_{i=1}^{U} \frac{e_{ij}}{e_i}$, $s_j = \frac{e_j}{e}$, $s_i = \frac{e_i}{e}$ and $HHI_U = \sum_{i=1}^{U} s_i^2 HHI_i$
Howard et al.		$CL_{ab} = \frac{\sum_{k=1}^{K_a} \sum_{l=1}^{K_b} C_{kl}}{K_a K_b}$ $XCL_{ab} = CL_{ab} - CL_{ab}^{RND}$ where: $C_{kl} = 1$ if firms k and l are located in the same region and $C_{kl} = 0$ otherwise

Notes: e_{ij} is the employment of industry i in region j, e_i is the total employment in industry i, e_j is the total employment in region j, e is the total employment in the whole economy, e_{ik} is the employment of firm k from industry i, k and l are indices for single firms, I is the number of industries, J is the number of regions, U is a subset of all I industries ($U \leq I$), K is the number of firms and K_a and K_b are the numbers of firms in industry a and b.

Compiled from: Farhauer, Kröll (2014); Howard et al. (2016); Nakamura, Morrison Paul (2009)

The most popular indicator is the Location Quotient (LQ), which is attributed to Hoover (1936) and mathematically equivalent to the Revealed Comparative Advantage (RCA) index, developed by Balassa (1965) in the context of international trade. The LQ is utilized in many studies (e.g. Bai et al. 2008; Kim 1995) as well as in the *OECD Territorial Reviews* (OECD, 2019). Following O'Donoghue, Gleave (2004) and Tian (2013), the original formulation can be extended: As the location quotient is not normalized, there is no cut-off value for defining a cluster, which leads to a standardization of the computed values via z-transformation. Hoen, Oosterhaven (2006) developed an additive alternative to the RCA index. The original LQ provides the main mathematical basis for several indicators developed later, such as the spatial Gini coefficients described below.

Some indicators which are known from the context of regional inequality (see Section 2) are also used for the analysis of agglomeration: A modification of the Gini coefficient is used for the spatial concentration of industries as well as regional specialization (e.g. Ceapraz 2008; Wieland, Fuchs 2018). As we can see in the calculation of R_i and C_j, respectively, the spatial Gini coefficient is based on the LQ. Another popular option for analyzing agglomeration is the Hoover coefficient, comparing the structure of an industry/a region to a reference structure of all industries/regions (e.g. Dixon, Freebairn 2009; Jiang et al. 2007). Both indicator types range between zero (no specialization/concentration) and one (total specialization/concentration). Also the Herfindahl-Hirschman index and its derivates are used to measure concentration, specialization and diversification (e.g. Duranton, Puga 2000; Goschin et al. 2009; Lehocký, Rusnák 2016).

Another type of specialization/concentration indicator was introduced by Krugman (1991), originally designed for comparing the specialization of *two* regions. An extension of this indicator was established by Midelfart-Knarvik et al. (2000) for the comparison of regional specialization/industry concentration with respect to the sum or mean of *all*

regions/industries (furthermore used e.g. by Haas, Südekum 2005; Vogiatzoglou 2006). Unlike the Gini- or Hoover-type measures, the Krugman coefficients range between zero (no specialization/concentration) and two (total specialization/concentration).

The cluster index developed by Litzenberger, Sternberg (2006) goes beyond employment data and includes additional information about the industry-specific firm size, population density and region size. It is composed of three parts: the relative industrial stock with respect to industry i and region j, IS_{ij}, the relative industrial density, ID_{ij}, and the relative firm size, PS_{ij}. All three components are modified location quotients. This is done to control for small and monostructural regions, which are identified as clusters otherwise (which is a problem in the original LQ). The cluster index CI_{ij} has a potential range from zero to infinity. This extended indicator is used e.g. by Hoffmann et al. (2017) for the German food processing industry.

The cluster indicators by Ellison, Glaeser (1997) compare the empirical distribution of firms to an arbitrary location pattern where agglomeration economies are absent (often referred to as a *dartboard approach*). Ellison, Glaeser (1997) differentiate between the clustering of firms from one industry (agglomeration) due to localization economies and the clustering of multiple industries (coagglomeration) due to urbanization economies. Their indices also take into account the industry-specific structure of the firms by including the Herfindahl-Hirschman index, HHI_i, for the employment concentration in industry i. This is the reason why individual firm-level data is required for the computation. The Herfindahl-Hirschman indicator is included to control the raw measures of spatial concentration, G_i and G, for firm employment concentration, which occurs especially when there are just a few firms with many employees. The Ellison-Glaeser (EG) index for agglomeration, γ_i, is designed for identifying the clustering of industry i, while the coagglomeration index, γ_c aims at the clustering of a set of U industries, where $U \leq I$. Values of γ equal to zero imply the absence of agglomeration economies, while values above zero indicate positive effects due to spatial clustering. When γ is negative, firm locations are less spatially concentrated than expected on condition of the dartboard approach, which indicates negative agglomeration economies. The EG index is used in several current regional economic studies (e.g. Dauth et al. 2015, 2018; Yamamura, Goto 2018).

In contrast, Howard et al. (2016) argue that agglomeration economies should not be analyzed regarding employment but the firms itself. Their colocation index, CL_{ab}, sums the colocation of K_i and K_q firms from two industries, i and q, controlling for all possible combinations. This colocation measure is compared to a counterfactual location structure constructed via bootstrapping; specifically the arithmetic mean of a number of (e.g. 50) random assignments of the regarded firms to the locations. The value of the resulting excess colocation index, XCL_{ab}, ranges between -1 and 1.

4.2 Application in REAT

4.2.1 REAT functions for regional specialization and industry concentration

Table 7 shows the REAT functions for agglomeration measures based on aggregate (employment) data. All functions require at least information about the employment in one or more regions j in one or more industries i, e_{ij}. The Herfindahl-Hirschman index (function `herf()`) for measuring regional diversity is not displayed as it is used exactly in the same way as described in Section 2, replacing x_i with e_{ij}.

Location quotients for one region and one or more industries are computed by the function `locq()`, including the option for an additive indicator instead of the multiplicative. When calculating the LQ for a set of J regions and I industries, one can use function `locq2()`, which is a kind of batch processing extension of `locq()`. As the dimension of the Litzenberger-Sternberg cluster index is the same as in the LQ (a single value for each combination of region j and industry i), the related functions `litzenberger()` and `litzenberger2()` work in the same way. When using `locq2()` or `litzenberger2()`, the user may choose the type of function output: either a `matrix` with I columns and J rows or a `data frame` with $I * J$ rows.

The Hoover-, Gini- and Krugman-type indicators require the same kind of input data. The `hoover()` function was already explained in Section 2, as it can be also

Table 7: REAT functions for regional specialization and industry concentration

Indicator	REAT function	Mandatory arguments	Optional arguments	Output
Hoover LQ/ Balassa RCA incl. extensions	locq()	vectors or single values of e_{ij} and e_i, single values of e_j and e	LQ method, plot	Single value or matrix with LQ_{ij}
	locq2()	vectors of e_{ij}, industry ID i and region ID j	normalization, output type, remove NAs	matrix or data frame with $I * J$ values of LQ_{ij}
Hoover specialization/ concentration	hoover() (see Section 2)	vectors of e_{ij} and reference vector e_i or e_j	remove NAs	value: H_j or H_i
Gini specialization	gini.spec()	vectors e_{ij} and e_i	plot LC	value: G_j, optional: LC plot
concentration	gini.conc()	vectors e_{ij} and e_j	plot LC	value: G_i, optional: LC plot
Krugman specialization	krugman.spec() (regions j and l)	vectors e_{ij} and e_{il}		value: K_{jl}
	krugman.conc2() (all J regions)	vector e_{ij} and matrix or data frame e_{il}		value: K_j
concentration	krugman.conc() (industries i and u)	vectors e_{ij} and e_{uj}		value: K_{iu}
	krugman.conc2() (all I industries)	vector e_{ij} and matrix or data frame e_{uj}		value: K_i
All at once: specialization	spec()	vectors of e_{ij}, industry ID i and region ID j	remove NAs	matrix with H_j, G_j and K_j (columns) for J regions (rows)
concentration	conc()	vectors of e_{ij}, industry ID i and region ID j	remove NAs	matrix with H_i, G_i and K_i (columns) for I industries (rows)
Duranton-Puga	durpug()	vectors e_{ij} and e_i		value: RDI_j
Litzenberger-Sternberg	litzenberger()	single values of e_ij, e_i, a_j, a, p_j, p, b_{ij} and b_i		value: CI_{ij}
	litzenberger2()	vectors of e_{ij}, industry ID i, region ID j, a_j, p_j and b_{ij}	output type, remove NAs	matrix or data frame with $I * J$ values of CI_{ij}

Source: own compilation.

used for measuring spatial concentration of industries or the specialization of regions with all-over employment vectors, e_i and e_j, respectively, as reference distributions. The spatial Gini coefficients are available through functions `gini.spec()` for regional specialization and `gini.conc()` for spatial concentration. The Krugman coefficients are divided into functions for the comparison of two regions/industries (`krugman.spec()` and `krugman.conc()`, respectively) and for applying all regions/industries as reference (`krugman.spec2()` and `krugman.conc2()`, respectively). The functions `spec()` and `conc()` are wrapper functions providing a convenient way to compute Hoover, Gini and Krugman coefficients of a given set of J regions and I industries at once, e.g. originating from official statistics on regional employment.

Table 8 shows the functions operating on the level of individual firm data. The Ellison-Glaeser (*EG*) indices are available through the functions `ellison.a()` (agglomeration index for industry i) and `ellison.a2()` (agglomeration indices for I industries) as well as `ellison.c()` (coagglomeration index for U industries) and `ellison.c2()` (coagglomeration indices for $I * I - I$ industry combinations). All functions require the firm size (e.g. no. of employees) for the k-th firm from industry i (`numeric vector`) and the

Table 8: REAT functions for agglomeration and coagglomeration using firm data

Indicator	REAT function	Mandatory arguments	Optional arguments	Output
Ellison-Glaeser agglomeration	ellison.a()	vectors of e_{ik}, e_j and region ID j		visible: value γ_i, invisible: matrix with γ_i, G_i, z_i, K_i and HHI_i
	ellison.a2()	vectors e_{ik}, industry ID i and region ID j		visible: values γ_i, invisible: matrix with γ_i, G_i, z_i, K_i and HHI_i, for I industries (rows)
coagglomeration	ellison.c()	vectors e_{ik}, industry ID i and region ID j	vectors e_j and U industries	value: γ^c
	ellison.c2()	vectors e_{ik}, industry ID i and region ID j	vector e_j	matrix with γ^c for $I * I - I$ industry combinations (rows)
Howard et al. colocation	howard.cl()	firm ID k, industry ID i, and region ID j, industries a and b		value: CL_{ab}
excess colocation	howard.xcl()	firm ID k, industry ID i and region ID j, industries a and b, no. of samples		value: XCL_{ab}
	howard.xcl2()	firm ID k, industry ID i and region ID j		matrix with XCL_{ab} for $I * I - I$ industry combinations (rows)

Source: own compilation.

region j the firm is located in. The functions incorporating more than one industry (all except for `ellison.a()`) require a `vector` containing the industry i. The data could e.g. be stored in a `data frame` with at least three columns (firm size, region, industry). Like some of the convergence functions (see Section 3), the EG agglomeration index functions in `REAT` also distinguish between a visible and an invisible output: `ellison.a()` and `ellison.a2()` show the value(s) auf γ_i but return an invisible `matrix` including the raw measure of concentration (G_i), the z-standardized results (z_i) and the related Herfindahl-Hirschman index for industry-specific firm concentration (HHI_i) as well as the number of firms in industry i (K_i).

The Howard-Newman-Tarp coagglomeration measure is distributed over the functions `howard.cl()` (calculation of the colocation index for one pair of industries a and b), `howard.xcl()` (calculation of the excess colocation index for industries a and b) and `howard.xcl2()` (calculation of the excess colocation index for $I * I - I$ combinations of I industries). As this cluster index works with firms instead of employment, we only need a `vector` containing the IDs of the firms k, the corresponding industry i and the region j where the firm is located. When calculating this measure for one pair of industries, the user must state the IDs of industries a and b. Note that calculation time for this index increases heavily with the number of firms and/or industries.

4.2.2 Application example 1: Regional specialization of Göttingen

We use the German classification of economic activities (WZ2008) on the level of 21 sections (A-U) for the classification of industries in the following examples (see Table 9).

Starting with a simple example, we analyze the regional specialization of Göttingen, a city with a population of about 134,000 in Niedersachsen, Germany. The example dataset `Goettingen`, which is included in `REAT`, contains the dependent employees in Göttingen and Germany for 2008 to 2017 in industries A to R (rows 2 to 16; row 1 contains the all-over employment). First, we load the data:

Table 9: Classification of economic activities in Germany, edition 2008 (WZ 2008)

WZ2008 Code	Title
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam and air conditioning supply
E	Water supply; sewerage, waste management and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
O	Public administration and defence; compulsory social security
P	Education
Q	Human health and social work activities
R	Arts, entertainment and recreation
S	Other service activities
T	Activities of households as employers; undifferentiated goods-and services-producing activities of households for own use
U	Activities of extraterritorial organisations and bodies

Source: own compilation based on Statistisches Bundesamt (2008).

```
data(Goettingen)
```

Using the `REAT` function `locq()`, we calculate a location quotient for Göttingen with respect to the manufacturing industry ("Verarbeitendes Gewerbe"), which is represented by letter C:

```
locq (Goettingen$Goettingen2017[4], Goettingen$Goettingen2017[1],
Goettingen$BRD2017[4], Goettingen$BRD2017[1])
# Industry: manufacturing (letter C) in row 4
# row 1 = all-over employment

[1] 0.5369
```

The output is simply the LQ value (LQ_{ij}, where i is manufacturing and j is Göttingen). We see that the LQ is very low, indicating that manufacturing is underrepresented in Göttingen as compared to Germany. Now, we calculate LQ values for all industries (A-R), including a simple plot (function argument `plot.results = TRUE`):

```
locq (Goettingen$Goettingen2017[2:16], Goettingen$Goettingen2017[1],
Goettingen$BRD2017[2:16], Goettingen$BRD2017[1],
industry.names = Goettingen$WZ2008_Code[2:16], plot.results = TRUE,
plot.title = "Location quotients for Göttingen 2017")
# all industries (rows 2-16 in the dataset)
```

The output is a `matrix` with one row for each industry:

```
Location quotients
I = 15 industries

            LQ
A   0.08407652
BDE 0.40085663
C   0.53687366
F   0.34366928
G   0.74603541
H   0.67117311
```

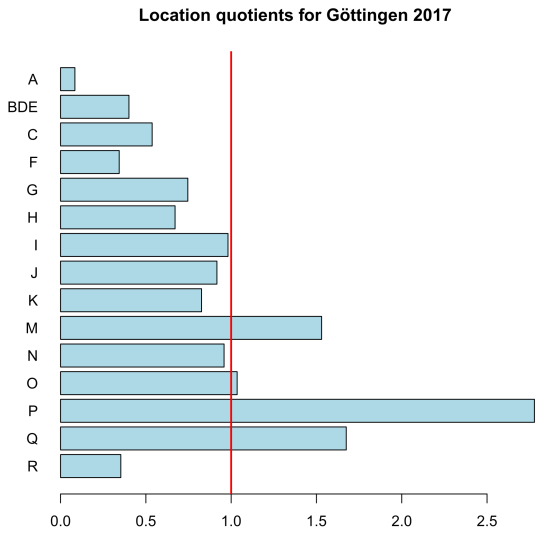


Figure 3: Location quotients for 15 industries in Göttingen

```
I   0.98141916
J   0.91654277
K   0.82650178
M   1.53027645
N   0.95843423
O   1.03509027
P   2.77790858
Q   1.67459967
R   0.35317012
```

The result is plotted in Figure 3. The function plots a vertical line at $LQ_{ij} = 1$ automatically. This is the (only) reference value for the LQ. It indicates a stock of the related industry equal to the whole economy. The highest LQ values can be found for the industries with letters P (education) and Q (health). This is because Göttingen is mainly characterized by a large university (about 30,000 students) with a university hospital with about 7,000 employees.

Now, we want to measure the specialization of Göttingen with a single indicator. First, we simply use the Herfindahl-Hirschman coefficient for both Göttingen and Germany using the function `herf()`:

```
herf(Goettingen$Goettingen2017[2:16])
[1] 0.127314

herf(Goettingen$BRD2017[2:16])
[1] 0.1104567
```

The *HHI* for Göttingen is slightly larger than for Germany, which indicates a higher specialization (or lower economic diversity) of the region. To combine this information in one indicator, we calculate the Hoover coefficient of specialization using the function `hoover()`, where the reference distribution is the German industry structure:

```
hoover(Goettingen$Goettingen2017[2:16], ref = Goettingen$BRD2017[2:16])
[1] 0.2254234
```

We finish our analysis of Göttingen's regional specialization by calculating both the Gini and the Krugman coefficient of regional specialization with the same data, using the `REAT` functions `gini.spec()` and `krugman.spec()`, respectively. Note that, here,

we use the Krugman coefficient to compare the industry structure of Göttingen to the structure of whole Germany (instead of another region within the country, for which this coefficient was originally formulated):

```
gini.spec(Goettingen$Goettingen2017[2:16], Goettingen$BRD2017[2:16])
[1] 0.359852

krugman.spec(Goettingen$Goettingen2017[2:16], Goettingen$BRD2017[2:16])
[1] 0.4508469
```

There seems to be some specialization in Göttingen, but, unfortunately, we do not have any real reference value to interpret the results.

4.2.3 Application example 2: Identifying clusters in Germany using aggregate data

In this example, we will compute indicators of regional specialization and industry concentration for a set of J regions and I industries at once. We load the included test dataset `G.regions.industries` containing employment and firms on the level of $I = 17$ industries (WZ2008 codes B-S) and $J = 16$ regions ("Bundesländer") in Germany:

```
data(G.regions.industries)
```

The number of employees in the column `emp_all` includes dependent employees and self-employed persons. The classification code of industries (see Table 9) can be found in column `ind_code`, while the region code (abbreviation of the region's official name) is in column `region_code`. First, we want to detect the spatial concentration of the 17 industries in Germany by calculating Hoover, Gini and Krugman coefficients for all industries at once, applying the `REAT` function `conc()` which is a wrapper function for the mentioned indicators. We save our output in the `matrix` object `conc_i`:

```
conc_i <- conc (e_ij = G.regions.industries$emp_all,
industry.id = G.regions.industries$ind_code,
region.id = G.regions.industries$region_code)
```

The output is:

```
Spatial concentration of industries
I = 17 industries, J = 16 regions

              H i        G i        K i
WZ08-B 0.22959050 0.42334831 0.45675385
WZ08-C 0.09933363 0.17047620 0.26813759
WZ08-D 0.07754576 0.12509360 0.16260016
WZ08-E 0.11972072 0.16742909 0.20369011
WZ08-F 0.07676634 0.15357575 0.16996098
WZ08-G 0.03034962 0.05471323 0.07977056
WZ08-H 0.06006957 0.11921850 0.10076748
WZ08-I 0.05177262 0.09939075 0.11450791
WZ08-J 0.10230712 0.22605802 0.24450967
WZ08-K 0.08982871 0.17610712 0.20565974
WZ08-L 0.09798632 0.16784764 0.17472656
WZ08-M 0.06490185 0.14760918 0.14931991
WZ08-N 0.06714816 0.08575299 0.09053327
WZ08-P 0.03019678 0.05053848 0.07043586
WZ08-Q 0.04679962 0.06170335 0.06406058
WZ08-R 0.09424708 0.16748405 0.17023603
WZ08-S 0.04507988 0.07246697 0.06441360
```

The function returns a `matrix` with 17 rows (one for each industry) and three columns: `H i` is the Hoover coefficient, `G i` is the Gini coefficient and `K i` is the Krugman coefficient for industry i. We cannot interpret or compare all of these results, but we may pick out some findings: The strongest spatial concentration is found with respect to mining

and quarrying (WZ08-B), no matter which indicator is regarded, which may be interpreted with "natural advantages" due to the spatial distribution of mineral resources in Germany. Services (such as retailing) as well as education and health are least concentrated, as these industries are bound to regional demand and/or their locations are regulated by policy and planning authorities.

At a first glance, the three indicators seem to produce similar results. Now, we want to test the similarity between Hoover, Gini and Krugman coefficients of concentration. As we saved our result `matrix`, we now calculate Pearson correlation coefficients (r) for each pair of indicators using the basic `R` function `cor()`, which is implemented in the `stats` package (included automatically in any `R` release). The function is applied to the three columns of `conc_i`, producing a $3 * 3$ correlation matrix:

```
cor(conc_i[,1:3])
          H i       G i       K i
H i 1.0000000 0.9676518 0.9527747
G i 0.9676518 1.0000000 0.9681770
K i 0.9527747 0.9681770 1.0000000
```

As we can see, each combination of the three indicators shows a strong positive correlation (H_i vs. G_i: $r \approx 0.97$, H_i vs. K_i: $r \approx 0.95$, G_i vs. K_i: $r \approx 0.97$). At least in this context, we may conclude that these indicators are interchangeable. However, we have to recognize that the analysis presented here is on a large-scale regional level (German "Bundesländer") and all of the mentioned indicators are affected by the *modifiable areal unit problem*, which means that the results depend on the aggregation unit in the analysis (see e.g. Dapena et al. 2016 for a discussion of this effect).

Now, we do exactly the same with respect to regional specialization of the 16 regions, using the same data. Analogously, we use the wrapper function `spec()` for calculating Hoover, Gini and Krugman coefficients of regional specialization, also saving the resulting `matrix`:

```
spec_j <- spec (e_ij = G.regions.industries$emp_all,
industry.id = G.regions.industries$ind_code,
region.id = G.regions.industries$region_code)
```

The output is:

```
Specialization of regions
I = 17 industries, J = 16 regions

          H j        G j        K j
BB 0.11530353 0.20632682 0.18555259
BE 0.17891265 0.29040841 0.34552331
BW 0.08024011 0.10300695 0.22675612
BY 0.05008135 0.07659148 0.16019603
HB 0.09502615 0.18563500 0.17467291
HE 0.05494422 0.12160142 0.11282696
HH 0.16413456 0.22616814 0.33190321
MV 0.13270849 0.18974606 0.22056868
NI 0.03772799 0.08237225 0.07972852
NW 0.02940091 0.05997505 0.07181569
RP 0.04793147 0.07432361 0.12036513
SH 0.08901907 0.11384295 0.15994524
SL 0.05726933 0.11921727 0.15071159
SN 0.05400855 0.10643512 0.10341280
ST 0.08821395 0.21120287 0.15280711
TH 0.08234046 0.13902924 0.17720208
```

The strongest specialization can be found in the city states Berlin (BE) and Hamburg (HH), while Niedersachsen (NI) and Nordrhein-Westfalen (NW) show the lowest values in all three indicators. As already mentioned in the concentration example, we have to remember the large-scale aggregation unit. If we used smaller scale units (e.g. counties like in Section 3.2.2), our results would surely be more differentiated. Again, we check the correlation between the indicators:

```
cor(spec_j[,1:3])
          H j       G j       K j
H j 1.0000000 0.9179127 0.9322604
G j 0.9179127 1.0000000 0.7907841
K j 0.9322604 0.7907841 1.0000000
```

Again, we find a strong positive correlation between the Hoover coefficient and both Gini and Krugman coefficient (H_j vs. G_j: $r \approx 0.92$, H_j vs. K_j: $r \approx 0.93$), while the third Pearson correlation coefficient is a little lower, but still showing the same direction (G_j vs. K_j: $r \approx 0.79$).

Now we check for clusters in a combination of a specific industry and a specific region. First, we calculate location quotients for the dataset `G.regions.industries` using the `REAT` function `locq2()`. Here, the optional function argument `LQ.norm` could be used for computing z-standardized location quotients according to O'Donoghue, Gleave (2004) (`LQ.norm = "OG"`) or z-standardized values of the natural-logged LQs according to Tian (2013) (`LQ.norm = "T"`). However, we produce the original LQs, since we need exactly the same columns as in the examples above:

```
locq2(e_ij = G.regions.industries$emp_all,
G.regions.industries$ind_code, G.regions.industries$region_code)
```

The output is a `matrix` with J rows and I columns:

```
Location quotients
I = 17 industries, J = 16 regions

              BB          BE        BW        BY        HB        HE
WZ08-B 2.5314363 0.04030901 0.6607950 0.8078054 0.0000000 0.3735773
WZ08-C 0.6857231 0.37224900 1.3968652 1.1902785 0.7863570 0.8580352
WZ08-D 1.1736475 0.46721079 1.0861988 0.8343784 0.9179718 0.9627955
WZ08-E 1.7945685 1.30128835 0.5896526 0.7137388 1.2393228 0.8532203
WZ08-F 1.5997778 0.77160121 0.9070096 1.0280409 0.6212923 0.8927681
WZ08-G 0.9550127 0.83133221 0.9492523 0.9879826 0.9013193 1.0006321
WZ08-H 1.3212794 0.87982228 0.8189666 0.8664163 1.7692815 1.2208728
WZ08-I 1.0379426 1.35561299 0.9132949 1.0390886 0.9904308 0.9571339
WZ08-J 0.5625876 1.78334039 1.0316114 1.1550764 1.0577107 1.1407078
WZ08-K 0.6529329 0.76630600 0.9329930 1.1058890 0.7825178 1.6710583
WZ08-L 1.1088846 2.13220960 0.7310014 0.8633894 1.3254723 1.1132939
WZ08-M 0.7366238 1.39880205 1.0337139 1.0265993 1.1202532 1.1457770
WZ08-N 1.2571301 1.24261162 0.7977054 0.8486971 1.2938161 1.0525912
WZ08-P 0.9052976 1.38842157 0.9649289 0.9252245 1.0563169 1.0085207
WZ08-Q 1.1540423 1.09329902 0.8695241 0.9079679 0.9544891 0.8980680
WZ08-R 1.0656945 2.55595102 0.8518192 0.8220540 1.3196613 0.8651451
WZ08-S 1.1409373 1.32596177 0.8626829 0.9092125 1.1396616 1.0528184
              HH        MV        NI        NW        RP        SH
WZ08-B 0.6029388 0.6235796 1.4987086 1.4595767 1.0371236 0.5078145
WZ08-C 0.4781934 0.6230156 0.9512438 0.9312325 1.0822678 0.7082513
WZ08-D 0.4332870 1.0118838 0.9932719 1.2139740 0.9349679 1.1248685
WZ08-E 1.1442005 1.5642257 1.0645497 1.0408356 0.9886860 1.0707585
WZ08-F 0.5432163 1.2716537 1.0969756 0.8735506 1.1134885 1.1043449
WZ08-G 1.0654315 0.9485377 1.0758977 1.0612190 1.0111274 1.2456100
WZ08-H 1.4958610 1.1243732 1.0409143 0.9961224 0.9633972 1.0112557
WZ08-I 1.0634066 1.7574637 1.0227196 0.8750205 1.1121264 1.2483966
WZ08-J 1.9266913 0.4751473 0.6716376 0.9830609 0.8122058 0.7496925
WZ08-K 1.5175078 0.5383900 0.9108456 1.0205798 0.8879292 0.8355178
WZ08-L 1.5871838 1.3034074 0.8040270 0.9928161 0.7774500 1.1980553
WZ08-M 1.6293913 0.6897571 0.8693026 1.0366589 0.7764558 0.7905498
WZ08-N 1.2530608 1.2484353 0.9675147 1.0659893 0.8026181 0.9727871
WZ08-P 0.9422739 0.9966228 1.0888054 0.9846351 1.0976178 0.9540262
WZ08-Q 0.8564604 1.2893168 1.0728412 1.0595648 1.0418460 1.1662290
WZ08-R 1.4914564 1.0500685 0.9204586 0.9611539 0.8498053 1.0418794
WZ08-S 0.8055128 1.1158184 0.9965451 1.0283571 1.1658852 1.1455178
```

```
              SL        SN        ST        TH
WZ08-B 0.2826284 1.2746172 2.4654331 0.7140637
WZ08-C 1.1752810 0.9867417 0.9297172 1.1849897
WZ08-D 1.1465539 1.0637093 1.2642787 0.8607578
WZ08-E 0.9555581 1.4457486 1.8251853 1.6042935
WZ08-F 0.9016858 1.3794286 1.4104724 1.3481005
WZ08-G 1.0370901 0.8787739 0.9172598 0.8661184
WZ08-H 0.8851047 1.0476688 1.2012430 0.8944907
WZ08-I 0.9111877 0.9496370 0.9020582 0.8644257
WZ08-J 0.7133587 0.7717704 0.4874344 0.6869177
WZ08-K 1.0082983 0.6620719 0.6133933 0.6316347
WZ08-L 0.7018816 1.1395422 1.0111694 0.8511896
WZ08-M 0.8060753 0.8459317 0.6627301 0.6814339
WZ08-N 1.0751749 1.1656467 1.2796548 1.1093251
WZ08-P 0.9147874 0.9658590 0.9798932 0.9710576
WZ08-Q 1.0760969 1.0475595 1.1401680 1.0628602
WZ08-R 0.7631263 1.1419135 0.8329295 0.8582919
WZ08-S 0.8840741 0.9774726 0.9257397 1.0923137
```

These $I * J = 17 * 16 = 272$ coefficients are too much information. Thus, we calculate them again using the optional argument `LQ.output = "df"`, which produces a `data frame` with $I * J$ rows and three columns (`j_region`: ID of region j, `i_industry`: ID of industry i and `LQ`: location quotient LQ_{ij}). We save the results in the object `lqs`:

```
lqs <- locq2(e_ij = G.regions.industries$emp_all,
G.regions.industries$ind_code, G.regions.industries$region_code,
LQ.output = "df")
```

As we forego an inspection of these singe values, the results are not displayed here. Instead, we only deal with the five highest LQs in our results (the "top five"). We sort the resulting `data frame` decreasing and take a look at the first five rows:

```
lqs_sort <- lqs[order(lqs$LQ, decreasing = TRUE),]
# Sort decreasing by size of LQ

lqs_sort[1:5,]

    j_region i_industry       LQ
33        BE     WZ08-R 2.555951
1         BB     WZ08-B 2.531436
239       ST     WZ08-B 2.465433
28        BE     WZ08-L 2.132210
111       HH     WZ08-J 1.926691
```

The highest LQ is found for the arts, entertainment, and recreation sector (WZ08-R) in the German capital Berlin. Note that this result is congruent with several studies about the "creative class", showing a large stock of "creative" employment in Berlin (e.g. Martin 2015). We also find a strong concentration of mining and quarrying in two Eastern regions, Brandenburg and Sachsen-Anhalt. Note that the LQ is a *relative* measure with respect to the total regional employment as well as the total industry-specific employment and the employment in the whole economy, not considering other aspects of industry or spatial structure.

These deficiencies should be overcome with the Litzenberger-Sternberg cluster index, also taking into account area, population and firm size. This additional data is also included in our current dataset (columns `area_sqkm`, `pop` and `firms`). The functions `litzenberger()` and `litzenberger2()` work equivalently to `locq()` and `locq2()`. To compute cluster indices for all $I * J$ combinations, we use the function `litzenberger2()`:

```
litzenberger2(G.regions.industries$emp_all,
G.regions.industries$ind_code, G.regions.industries$region_code,
G.regions.industries$area_sqkm, G.regions.industries$pop,
G.regions.industries$firms)
```

Like in `locq2()`, the default output is a `matrix` with I rows and J columns:

```
Litzenberger-Sternberg cluster indices
I = 17 industries, J = 16 regions

              BB           BE        BW        BY        HB        HE
WZ08-B 0.5736692   0.05611505 0.8041813 1.1073446       NaN 0.4745084
WZ08-C 0.1610679   3.24717820 2.6250805 1.2043415  5.119669 1.0603087
WZ08-D 0.2213627   1.37720778 1.7043505 1.4178208  4.172162 0.8541359
WZ08-E 0.8810260  10.14891585 0.8235517 0.6890213  6.705744 1.1427285
WZ08-F 0.7142888  11.36434353 1.2372108 0.9442221  3.498921 1.1225087
WZ08-G 0.2707787  12.10626625 1.3903532 0.9404677  7.136205 1.3361060
WZ08-H 0.4386878  13.26265081 1.0955747 0.7982074 22.656924 1.8358272
WZ08-I 0.2672336  26.60020727 1.4098657 0.9633029  8.481338 1.2880210
WZ08-J 0.1130326  59.24931837 1.4342037 1.2393579  8.683998 1.9024826
WZ08-K 0.1524825  10.28664194 1.4774980 1.1692461  6.650213 2.4739044
WZ08-L 0.2564814  56.65943460 0.9594093 0.8695636 11.839900 1.6099929
WZ08-M 0.1685895  39.30149403 1.5306799 1.0110472  9.410498 1.7302434
WZ08-N 0.4232166  26.91532975 1.0228326 0.7471872 10.796027 1.5265846
WZ08-P 0.2043023  25.30839556 1.3656409 0.9509028  7.322709 1.4627871
WZ08-Q 0.3445630  21.86956483 1.1770297 0.7873877  7.850886 1.2163624
WZ08-R 0.2450932 104.73565741 1.0839767 0.7821779 10.555369 1.0489672
WZ08-S 0.2891132  24.71310833 1.3435576 0.9594882  9.893688 1.5421903
              HH         MV        NI        NW        RP        SH
WZ08-B  2.319611 0.16000177 1.5004530 2.074735 1.1951266 0.3119091
WZ08-C  4.000104 0.11993714 0.5036838 2.010757 0.9646351 0.4008598
WZ08-D  1.679371 0.20954802 0.8848956 2.001708 0.6016715 1.2989105
WZ08-E 11.129156 0.44055556 0.6476797 1.915196 0.8616891 0.8101087
WZ08-F  5.526083 0.45291220 0.6276791 1.656384 0.8973162 0.8055662
WZ08-G 15.627090 0.22665565 0.6898359 2.377365 0.8115817 0.9008287
WZ08-H 44.371836 0.32192237 0.6420146 1.980491 0.7087341 0.7205961
WZ08-I 14.795885 0.59842705 0.6335126 1.731963 1.0572996 1.0361389
WZ08-J 59.720584 0.05895184 0.2905769 2.139796 0.5142953 0.4427180
WZ08-K 26.189623 0.12112900 0.5648050 1.953963 0.7104929 0.5758818
WZ08-L 33.175443 0.25167830 0.5228248 2.223854 0.5349641 0.8587588
WZ08-M 44.433527 0.11657637 0.4440959 2.297308 0.4961611 0.4456119
WZ08-N 23.255195 0.31231467 0.5271293 2.413351 0.5537510 0.7348068
WZ08-P 14.294075 0.24008982 0.7203953 2.022971 0.9704975 0.6937792
WZ08-Q 13.211918 0.38626491 0.6877487 2.260518 0.7770980 0.8426253
WZ08-R 44.256214 0.20066782 0.4508862 2.190752 0.5287290 0.6694273
WZ08-S 14.195156 0.27631480 0.5364848 1.941470 0.7996291 0.9238816
              SL        SN         ST        TH
WZ08-B 0.3673090 1.1179110 1.49064630 0.4562730
WZ08-C 1.8348713 1.0311722 0.32128528 0.7892360
WZ08-D 0.9643263 0.4534029 0.29391783 0.2086042
WZ08-E 1.9939461 1.5554125 1.09973945 1.1815726
WZ08-F 1.3245152 1.8167263 0.68075807 1.0313749
WZ08-G 1.7528078 0.7644695 0.31005922 0.4313306
WZ08-H 1.1099151 0.9297053 0.42973342 0.4881643
WZ08-I 1.7871163 0.7119567 0.29998391 0.4012910
WZ08-J 0.8984679 0.4927030 0.08046178 0.2087174
WZ08-K 1.5505928 0.5980588 0.21942805 0.3160432
WZ08-L 0.8170723 0.8226773 0.21044244 0.2886068
WZ08-M 1.0137151 0.6489868 0.15797219 0.2493458
WZ08-N 1.3940298 1.2492658 0.42285503 0.5935329
WZ08-P 1.2331390 0.7800531 0.31406013 0.4389675
WZ08-Q 1.8551266 1.0749000 0.48353914 0.5854787
WZ08-R 0.8477702 0.8666478 0.21562997 0.2919300
WZ08-S 1.6138955 0.8600519 0.34624522 0.5369766
```

Note that there is a value equal to `NaN`, which means "not a number", due to a division by zero; this is because there is no mining and quarrying (WZ08-B) in Bremen (HB). However, we take a look at the "top five" again:

```
lss <- litzenberger2(G.regions.industries$emp_all,
G.regions.industries$ind_code, G.regions.industries$region_code,
G.regions.industries$area_sqkm, G.regions.industries$pop,
G.regions.industries$firms, CI.output = "df")

lss_sort <- lss[order(lss$CI, decreasing = TRUE),]

lss_sort[1:5,]

    j_region i_industry        CI
33        BE     WZ08-R 104.73566
111       HH     WZ08-J  59.72058
26        BE     WZ08-J  59.24932
28        BE     WZ08-L  56.65943
114       HH     WZ08-M  44.43353
```

Again, we find the largest cluster value for the arts and entertainment sector in Berlin. Also the other four highest indicators are discovered in the largest city states Berlin and Hamburg, especially with respect to the information and communication industry (WZ08-J) and other knowledge-intensive services. Obviously, the results of the Litzenberger-Sternberg index differ in a noticeable way from those of the LQ, which can be attributed to the consideration of other spatial aspects, especially controlling for the size of the regions.

4.2.4 Application example 3: Identifying clusters using micro-data

In our last example about agglomerations, we use the Ellison-Glaeser indices and the Howard-Newman-Tarp colocation index, which both require individual firm data. As this kind of micro-data is sensitive and, of course, not available in official statistics, we have to use fictional data from the textbook by Farhauer, Kröll (2014).

At first, we compute the Ellison-Glaeser agglomeration index for one industry i, γ_i. We use the `REAT` function `ellison.a()`, which is designed for this purpose and requires three `vectors`: the size (employment) of firm k, e_{ik}, the IDs of the regions j each firm is located in, and the total regional employment, e_j. The numerical example in Farhauer, Kröll (2014), Table 14.11, contains ten firms in three regions (Wien, Linz, and Graz). We simply compile the data from the original table into separate `vectors`:

```
region <- c("Wien", "Wien", "Wien", "Wien", "Wien", "Linz",
"Linz", "Linz", "Linz", "Graz")
# regions (Austrian cities)
emp_firm <- c(200,650,12000,100,50,16000,13000,1500,1500,25000)
# employment of the ten firms
emp_region <- c(500000,400000,100000)
# employment of the three regions
```

Now, we apply `ellison.a()` to this data:

```
ellison.a (emp_firm, emp_region, region)
[1] 0.05990628
```

The EG agglomeration index of $\gamma_i \approx 0.06$, which is, by the way, the same result as in the textbook, indicates a stronger clustering than expected from a dartboard approach. Since this data is fictional, we refrain from interpreting this result.

The `REAT` package contains the dataset `FK2014_EGC`, which is compiled from the numerical example in Farhauer, Kröll (2014), Tables 14.14 to 14.17. There are $k = 42$ firms from $I = 4$ industries (clothing trade, forestry, textiles dyeing and textiles trade) in $J = 3$ regions (1, 2 and 3). We load this example data:

```
data(FK2014_EGC)
```

We compute γ_i for all industries in the dataset. This can be done with the function `ellison.a2()`, which requires `vectors` containing the size of firm k, the corresponding industry i, and region j. We save the results in the object `ega`:

```
ega <- ellison.a2 (FK2014_EGC$emp_firm, FK2014_EGC$industry,
FK2014_EGC$region)
```

Here, we see the output of the function:

```
Ellison-Glaeser Agglomeration Index
K = 42 firms, I = 4 industries, J = 3 regions

                    Gamma i
Clothing trade  -0.09379384
Forestry         0.16838003
Textiles dyeing -0.08012539
Textiles trade  -0.13040134
```

We see a strong clustering of the forestry industry, which is attributed to localization economies, but spatial avoidance in the three other industries. The visible output of `ellison.a2()` contains the γ_i values only, but the invisible `matrix` output also includes the other information referring to the *EG* agglomeration index:

```
ega
                    Gamma i         G i       z i K i      HHI i
Clothing trade  -0.09379384 0.017909653 -0.5025978  11 0.13124350
Forestry         0.16838003 0.088262934  1.3660878  13 0.09240553
Textiles dyeing -0.08012539 0.027764811 -0.3801644   9 0.14559983
Textiles trade  -0.13040134 0.002734966 -0.7663541   9 0.12208059
```

When looking at the forestry industry, we also see a high standardized value ($z_i \approx 1.37$) and a relatively low firm concentration ($HHI_i \approx 0.09$).

In the next step, we compute the *EG* coagglomeration index, γ^c, for the same data using the function `ellison.c()`. This function requires the same information as `ellison.a2()` plus the total employment in the regarded regions (column `emp_region`):

```
ellison.c (FK2014_EGC$emp_firm, FK2014_EGC$industry,
FK2014_EGC$region, FK2014_EGC$emp_region)

[1] 12.0729
```

Congruent with the calculation in Farhauer, Kröll (2014), the function returns $\gamma^c \approx 12.07$. This value is very large, which indicates urbanization economies in this fictional example.

If we want to analyze the coagglomeration of industry pairs instead, we may use the function `ellison.c2()`, which requires the same data:

```
ellison.c2 (FK2014_EGC$emp_firm, FK2014_EGC$industry,
FK2014_EGC$region, FK2014_EGC$emp_region)
```

The output is a `matrix` with $I * I - I$ rows (one for each industry pair, omitting the combination of the same industry i):

```
Ellison-Glaeser Co-Agglomeration Index
K = 42 firms, I = 4 industries, J = 3 regions

                                Gamma c
Forestry-Clothing trade        1.382257
Textiles dyeing-Clothing trade 2.465609
Textiles trade-Clothing trade  2.067766
Clothing trade-Forestry        1.382257
Textiles dyeing-Forestry       1.570292
Textiles trade-Forestry        1.336020
Clothing trade-Textiles dyeing 2.465609
Forestry-Textiles dyeing       1.570292
Textiles trade-Textiles dyeing 2.294259
Clothing trade-Textiles trade  2.067766
Forestry-Textiles trade        1.336020
Textiles dyeing-Textiles trade 2.294259
```

If we want to focus on firm numbers instead of employment size, we may compute the Howard-Newman-Tarp excess colocation index, which is included in `REAT` through the functions `howard.cl()` for one colocation index for one pair of industries, `howard.xcl()` for the corresponding excess colocation index and `howard.xcl2()` for all combinations of $I * I$ industries. Subsequent to the numerical example above, we calculate XCL_{ab} for all industry pairs in the dataset `FK2014_EGC`, where the firm ID of k is stored in the column `firm`:

```
howard.xcl2 (FK2014_EGC$firm, FK2014_EGC$industry,
FK2014_EGC$region)
# this takes some seconds
```

The output has the same structure as the output from `ellison.c2()`:

```
Howard-Newman-Tarp Excess Colocation Index
K = 42 firms, I = 4 industries, J = 3 regions

                                    XCL
Forestry-Clothing trade        0.01902098
Textiles dyeing-Clothing trade 0.02909091
Textiles trade-Clothing trade  0.02020202
Clothing trade-Forestry        0.02377622
Textiles dyeing-Forestry       0.03282051
Textiles trade-Forestry        0.03589744
Clothing trade-Textiles dyeing 0.02707071
Forestry-Textiles dyeing       0.02666667
Textiles trade-Textiles dyeing 0.02814815
Clothing trade-Textiles trade  0.02101010
Forestry-Textiles trade        0.01743590
Textiles dyeing-Textiles trade 0.03012346
```

We see that the index by Howard et al. (2016) is structured differently than the indicators presented above: Although they are based on exactly the same data, the value for forestry and clothing trade ($XCL \approx 0.019$) is *not* equal to the value for clothing trade and forestry ($XCL \approx 0.024$). Why? The XCL_{ab} is the difference between the colocation index, CL_{ab}, and the mean of a set of bootstrap samples, CL_{ab}^{RND} (see Table 6). These random samples are drawn again each time a XCL value is computed, consequently, also the XCL value changes.

5 Proximity and accessibility

5.1 Distance-based measures of accessibility and proximity using individual point-level data

In this chapter, we mix two different concepts of indicators, accessibility and spatial proximity (see Table 10), both frequently used especially in the context of GIS (geographic information systems). Both concepts are discussed together because they have two aspects in common: 1) they are based on the geographical distance between point locations, in particular, the distance between an origin point i or several origin points ($i = 1, ..., n$) and one or more destination points j ($j = 1, ..., m$), and 2) for the calculation, they require geocoded (with geographical coordinates) individual point data.

One popular indicator of accessibility is the Hansen accessibility, developed by Hansen (1959) in the context of land use theory. The basic idea is that "accessibility" equals the sum of opportunities outgoing from a specific origin i. These opportunities are spread over a set of m locations ($j = 1, ..., m$). The summation is weighted with the distance between i and the j-th location. This distance, no matter how measured (e.g. street distance, Euclidean distance, driving time) is assumed to be perceived in a nonlinear way, which is operationalized by a nonlinear distance decay function (a.k.a. distance impedance function or response function), e.g. power, exponential or logistic. A similar concept was introduced by Harris (1954) attempting to model the market potential of

Table 10: Accessibility and proximity indicators using point-level data

Indicator	Non-normalized	Normalized	
Accessibility/Market potential			
Harris	$M_j = \sum_{i=1}^{n} O_i d_{ij}^{-1}$ $0 \leq M_j \leq \infty$		
Hansen	$A_i = \sum_{j=1}^{m} O_j f(d_{ij})$ $i \neq j$ $0 \leq A_i \leq \infty$	$A_i^* = \frac{\sum_{j=1}^{m} O_j f(d_{ij})}{\sum_{j=1}^{m} O_j}$ $i \neq j$ $0 \leq A_i^* \leq 1$	
	where: $f(d_{ij}) = d_{ij}^{-\lambda}$ or $f(d_{ij}) = e^{-\lambda * d_{ij}}$ or $f(d_{ij}) = \frac{1}{1+e^{-\lambda_1+\lambda_2 d_{ij}}}$		
Proximity			
Count within buffer	$N_i = \sum_{i=1}^{n} I(d_{ij} \leq t)$ $i \neq j$		
Weighted count within buffer	$N_i^w = \sum_{i=1}^{n} I(d_{ij} \leq t) O_j$ $i \neq j$		
Ripley	$K_t = \frac{1}{\lambda} \sum_{i=1}^{n} \frac{I(d_{ij} \leq t)}{n}$ $i \neq j$	$L_t = \sqrt{\frac{K_t}{\pi}}$ $i \neq j$	$H_t = L_t - t$ $i \neq j$
	$E(K_t) = \pi t^2$	$E(L_t) = t$	$E(H_t) = 0$
	where: $\lambda = \frac{n}{A}$		

Notes: d_{ij} is the distance from origin location i ($i = 1, ..., n$) to destination location j ($j = 1, ..., m$), O_j is a variable quantifying the size of destination j, t is a maximum search radius and $I(d_{ij} \leq t)$ is the indicator function taking the value of $I = 1$ if $d_{ij} \leq t$, and $I = 0$ otherwise.
Compiled from: Kiskowski et al. (2009); Krider, Putler (2013); Peña Carrera (2002); Pooler (1987); Reggiani et al. (2011); Smith (2016)

locations. If we replace the inverse distance weighting in the Harris indicator with another type of distance weighting, we see that both concepts are mathematically equivalent. The only difference is that the Harris indicator is conceptualized from the supplier's perspective j (e.g. market potential of a retail store) and the Hansen accessibility takes the demand location i as a starting point (Pooler, 1987; Reggiani et al., 2011). As these indicators are dimensionless and range from zero to infinity, a normalization with a range from zero to one can be computed by weighting the results with the opportunities without distance correction.

This accessibility/potential concept can be used in the regional economic context e.g. to quantify the over-regional job potential (e.g. Wieland, Fuchs 2018) or the clustering of point locations of a specific type, such as retail stores (e.g. Larsson, Öner 2014). The most common application of these indicators may be the context of transport economics and transport geography (e.g. Albacete et al. 2017).

In the GIS context, spatial proximity can be measured using concentric zones within a radius of t (buffers) around point i, where the number of the j points within this radius is counted (Longley et al., 2005). A systematic analysis of spatial proximity or cluster patterns is possible using Ripley's K function (Ripley, 1976). It compares empirical point counts with expected values from a random spatial point process based on a Poisson distribution. Ripley's K computes empirical values for each distance band with a maximum distance of t, which can be compared to the expected value. A more comprehensible (and linear) interpretation is provided when normalizing the K function in the form of the L or H function. Also, confidence intervals for the expected values can be calculated by bootstrapping (Kiskowski et al., 2009; Smith, 2016). All of these measures are based on a simple indicator function, $I(d_{ij} \leq t)$, which takes the value of $I = 1$ if point j is within a distance of t from point i or not ($I = 0$). Originating from natural sciences, especially Ripley's K is frequently used when analyzing location patterns in spatial economic contexts, such as the clustering of retail stores (e.g. Krider,

Table 11: REAT functions for accessibility and proximity on the point level

Indicator	REAT function	Mandatory arguments	Optional arguments	Output
Distance matrix	dist.mat()	data frame(s) with start points i (ID, lat, lon) and end points j (ID, lat, lon), distance unit	$i \neq j$	data frame with from, to, from-to and distance d_{ij} (distance matrix)
Buffer	dist.buf()	data frame(s) with start points (ID, lat, lon) and end points (ID, lat, lon), max. distance t, distance unit	$i \neq j$, sum O_j at endpoints	list with distance matrix (data frame) and count table (data frame)
Hansen/ Harris	hansen()	distance matrix (data frame with start points i and end points j as well as distance d_{ij} and O_j), weighting functions, parameters λ and γ	distance constant, max. distance t, $i \neq j$	data frame with origins i and accessibility A_i
Ripley	ripley()	data frame with points (ID, lat, lon), total area A, max. distance t, number of distance intervals	local K values, confidence intervals no. of samples, significance level, plot (K, L or H)	visible: matrix with t, K_t, $E(K_t)$, $K_t - E(K_t)$, L_t and H_t for each distance interval, invisible: matrix (as described above) and optional: matrices with local K values and confidence intervals

Source: own compilation.

Putler 2013) or other types of firms assumed to be connected in a network (e.g. Espa et al. 2010).

5.2 Application in REAT

5.2.1 REAT functions for accessibility and proximity on the point level

Table 11 shows the `REAT` functions for the accessibility and proximity methods described above. A simple Euclidean distance matrix for georeferenced points (`data frame` with latitude and longitude) can be calculated using the function `dist.mat()`. The function `dist.buf()` computes a "count points within buffer", where also a weighting, O_j, can be summarized (e.g., if the destination points are cities of a given population, one could count the number of cities within 50 kilometers and their corresponding population). The latter function uses `dist.mat()`, thus, it is not necessary to create a distance matrix before.

The same is the case for the function `ripley()`, which calculates Ripley's K function for georeferenced data (`data frame` with lat/lon) and a given number of distance intervals up to a maximum distance of t. The differences between the empirical values, K_t, and the expected values, $E(K_t)$, as well as the normalizations (L_t and H_t) are calculated and returned automatically. Optionally, local K values for each distance interval and corresponding confidence intervals are computed. These confidence intervals are based on bootstrapping with a given number of samples (default: 100) on a given significance level (the default value is $\alpha = 0.05$, which leads to confidence intervals of a range from $\alpha/2 = 2.5\%$ to $1-\alpha/2 = 97.5\%$). Note that the plot of the K function (or, when desired, L or H function) provides a graphical and more intuitive interpretation of the analyzed point pattern, especially when including confidence intervals.

When calculating the Hansen accessibility (or the Harris market potential) with `hansen()`, a distance matrix including the opportunities, O_j, is required. This can be, of course, done with `dist.mat()` (if straight-line distances are sufficient), but also with any other software creating distance matrices (and any type of transport costs indicator). In `hansen()`, the user may choose between a power, exponential or logistic distance decay function. Optionally, the normalized Hansen accessibility is returned additionally.

5.2.2 Application example 1: Location analysis of medical practices

In the example in Section 2.2.2, we dealt with small-scale regional inequality in health care in South Lower Saxony, Germany. We have seen that e.g. psychotherapists are more spatially clustered than general practitioners (GPs). Returning to this topic, we want to use proximity and accessibility measures for determining the market potential (in the sense of the Harris model) of these health care locations. Obviously, there are different location patterns of general practitioners and psychotherapists. In the related study, there was evidence that psychotherapists are not just clustered but clustered within some districts of larger cities (Wieland, Dittrich, 2016). In the German health care planning system, the market potential of medical practices is the main determinant of the official authorization to be included into the allocation system of health insurance, while psychotherapists are assumed to need quite larger market areas than GPs (Kassenärztliche Bundesvereinigung, 2013). Consequently, our research hypothesis is that the population potential of psychotherapists is larger than that of general practitioners.

We use the same test data as in the mentioned example, containing the health locations (`GoettingenHealth1`) and the corresponding settlements (`GoettingenHealth2`). We load both `R` datasets:

```
data(GoettingenHealth1)
data(GoettingenHealth2)
```

Table `GoettingenHealth1` contains 617 locations, whose ID is stored in the column `location`. Columns `lat` and `lon` contain the latitude and longitude, respectively, while the corresponding location type can be found in column `type` (`phys_gen`: general practitioners, `psych`: psychotherapists, `pharm`: pharmacies). As the following applications may be time-consuming, we extract the general practitioners from `GoettingenHealth1` and draw a random sample of ten doctor's practices:

```
physgen <- GoettingenHealth1[GoettingenHealth1$type == "phys_gen",]
# general practitioners: column "type" is equal to "phys_gen"
physgen_sample <- physgen[sample(nrow(physgen),10),]
# random sampling of ten general practitioners
```

Now, we want to summarize the population potential of these health locations in a 1,000 meters buffer. We apply the function `dist.buf()` to the sample data `physgen_sample` and sum up the local population of the districts within this distance (column `pop` in `GoettingenHealth2`):

```
physgen_pot <- dist.buf (physgen_sample, "location", "lat", "lon",
GoettingenHealth2, "district", "lat", "lon", bufdist = 1000,
ep_sum = "pop")
# counting all districts within a radius of 1000 meters
# and summing the corresponding population
```

We calculate the arithmetic mean of all ten potentials:

```
mean2(physgen_pot$count_table$sum_pop)
[1] 8027.7
```

On average, the ten GP practices have a population potential of about 8,028 inhabitants. One problem related to the buffer technique is the lack of distance weighting: All origin points up to a given distance are included completely, while all points above 1,000 meters are ignored. Thus, we repeat estimating the population potential using the Hansen accessibility. At first, we need an origin-destination matrix (distance matrix) from the origin points to the sampled GP locations. We use the function `dist.mat()` and merge the returned distance matrix with the population values from `GoettingenHealth2`:

```
physgen_od <- dist.mat(GoettingenHealth2, "district", "lat", "lon",
physgen_sample, "location", "lat", "lon")
# creating OD matrix from all districts to the
# sampled general practitioners

physgen_od <- merge (physgen_od, GoettingenHealth2,
by.x = "from", by.y = "district")
# merging with GoettingenHealth2 to include the
# population values of the districts
```

Then, we use the function `hansen()` to calculate the Hansen accessibility (used in the sense of the Harris market potential model) for each GP location in `physgen_od`.

The required columns in this dataset are the IDs of the GP locations (`to`), the IDs of the districts (`from`) and the population of the districts (`pop`) as well as the distances calculated above (`distance`). Finally, we have to set a distance weighting (which has an important influence in all types of spatial interaction models like this). For this purpose, we fall back on the results of a study by Fülöp et al. (2011): Based on empirical patient's choice of doctor, they estimated distance decay functions in spatial interaction models (Huff model) for several types of physicians. For GPs, an exponential distance decay function with $\lambda = -0.28$ was found to fit the empirical data best. To set a distance decay function type and the related weighting(s), the function arguments `dtype` and `lambda` must be used. We save the results under the name `physgen_hansen`:

```
physgen_hansen <- hansen (physgen_od, "to", "from", "pop",
"distance", dtype = "exp", lambda = -0.28)
# calculating Hansen accessibility for the ten
# sampled general practitioners
```

The output of the `hansen()` function is:

```
Hansen Accessibility

J = 420 locations with mean attractivity = 1138.486
I = 10 origins with mean transport costs = 28.07581
Attractivity weighting (pow) with Gamma = 1
Distance weighting (exp) with Lambda = -0.28

     to accessibility
1  1103     24267.054
2  1171     17629.564
3  1206      9581.732
4  1220      9213.407
5   197     10023.854
6   301      6489.571
7   600     69676.232
8   755     66921.123
9   966     13154.921
10  974      3666.171
```

Again, we calculate the arithmetic mean of the distance-weighted market potentials:

```
mean2(physgen_hansen$accessibility)
[1] 23062.36
```

The average population potential of the ten GPs is equal to 23,063 inhabitants.

As we want to compare the market potential of GPs and psychotherapists, we repeat the same analysis for them, now in the "fast mode", leaving out most comments, as the functions and commands are exactly the same as above, only applied to psychotherapists.

```
psychgen <- GoettingenHealth1[GoettingenHealth1$type == "psych",]

psych_sample <- psychgen[sample(nrow(psychgen),10),]

psych_pot <- dist.buf (psych_sample, "location", "lat", "lon",
GoettingenHealth2, "district", "lat", "lon", bufdist = 1000,
ep_sum = "pop")

mean2(psych_pot$count_table$sum_pop)
[1] 12245.88
```

The calculation of Hansen accessibility is different from the one for GPs with respect to the assumed distance reaction of the (potential) clients: For psychotherapists, Fülöp et al. (2011) found a distance impedance which is considerably smaller than for GPs (and any other type of doctor), resulting in a weighting parameter of $\lambda = -0.11$ in the exponential decay function:

```
psych_od <- dist.mat(GoettingenHealth2, "district", "lat", "lon",
psych_sample, "location", "lat", "lon")

psych_od <- merge (psych_od, GoettingenHealth2,
by.x = "from", by.y = "district")

psych_hansen <- hansen (psych_od, "to", "from", "pop",
"distance", dtype = "exp", lambda = -0.11)


Hansen Accessibility

J = 420 locations with mean attractivity = 1138.486
I = 10 origins with mean transport costs = 25.56756
Attractivity weighting (pow) with Gamma = 1
Distance weighting (exp) with Lambda = -0.11

     to accessibility
1  1031      43415.63
2  1213      39226.26
3   179      33491.41
4   313      51228.41
5   506     147887.43
6   786     147969.39
7   791     147971.80
8   811     148021.51
9   872     147475.57
10  922      42424.51

mean2(psych_hansen$accessibility)
[1] 94911.19
```

We see that the average population potential of the sampled psychotherapists on the 1,000 meters buffer level is equal to 12,246 inhabitants, which is about one third more than for GPs. The Hansen/Harris market potential of psychotherapists of about 94,911 persons is a fourfold increase compared to the GPs. We have to remember that the last result is not only a matter of location but also due to a lower assumed distance decay. However, the population potential of the sampled psychotherapists is obviously higher than the potential of the GPs, which can be attributed to a different location pattern, where psychotherapists are more clustered within larger city districts.

5.2.3 Application example 2: Clustering of health service providers

We stick to the example of health care locations. As we have found different degrees of regional inequality with respect to suppliers (Section 2.2.2) and of market potentials

(Section 5.2.2), we now analyze the clustering patterns of health service providers. In South Lower Saxony there is nearly the same number of psychotherapists (118) and pharmacies (120), but we should not expect their location patterns to be similar or even equal. Following the results above, we hypothesize that psychotherapists are more spatially clustered than pharmacies (as we already know about clustering with respect to districts in the former case and we can expect an avoidance tendency in the latter case due to a high degree of substitutability).

For this analysis, we compute Ripley's K with the `REAT` function `ripley()`. Before going on, we have to prepare two things: First, we load the required dataset. Then, we must calculate the total area of the study area manually (here: in square meters).

```
data (GoettingenHealth1)

area_goe <- 1753000000
# area of Landkreis Goettingen (sqm)
area_nom <- 1267000000
# area of Landkreis Northeim (sqm)
area_gn <- area_goe+area_nom
```

Now, we compute Ripley's K for the pharmacies only, which means processing only those locations in `GoettingenHealth1` which are pharmacies (`type == "pharm"`). We set our maximum search radius equal to $t = 30000$ (function argument `t.max`), divided into 300 distance intervals (`t.sep`), resulting in distance steps of 100 meters. As we want to check for a significant deviation from a random spatial pattern, we instruct the function to construct confidence intervals (`ci.boot = TRUE`) using the default settings ($\alpha = 0.05$, 100 bootstrapping samples). We also plot the results (default function argument: `K.plot = TRUE`) to inspect our results graphically. Here, we plot K_t, which is also the default setting (if the user wants to plot L_t or H_t instead, the function argument `Kplot.func` has to be changed to "`L`" or "`H`", respectively):

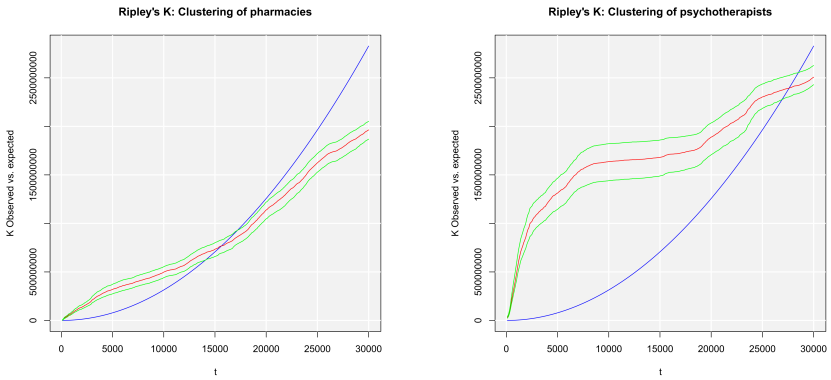
```
ripley(GoettingenHealth1[GoettingenHealth1$type == "pharm",],
"location", "lat", "lon", area = area_gn, t.max = 30000, t.sep = 300,
K.local = TRUE, ci.boot = TRUE, ci.alpha = 0.05, ciboot.samples = 100,
plot.title = "Ripley's K: Clustering of pharmacies")
```

The output is a `matrix` with six columns and one row for each distance interval. Thus, we skip the full output here:

```
Ripley's K
n = 120 points

     t <=       K t exp        K t  Kt-Kt exp       L t         H t
1     100      31415.93    3355556    3324140  1033.492   933.49238
2     200     125663.71   12583333   12457670  2001.349  1801.34940
3     300     282743.34   25586111   25303368  2853.824  2553.82412
4     400     502654.82   32297222   31794567  3206.326  2806.32580
5     500     785398.16   39008333   38222935  3523.739  3023.73923
...
```

We repeat the computation of Ripley's K for the psychotherapists:

```
ripley(GoettingenHealth1[GoettingenHealth1$type == "psych",],
"location", "lat", "lon", area = area_gn, t.max = 30000, t.sep = 300,
K.local = TRUE, ci.boot = TRUE, ci.alpha = 0.05, ciboot.samples = 100,
plot.title = "Ripley's K: Clustering of psychotherapists")
```



(a) Pharmacies (n=120) (b) Psychotherapists (n=118)

Figure 4: Plots of the Ripley-K function with confidence intervals

The output is (also truncated):

```
Ripley's K
n = 118 points

     t <=       K t exp         K t    Kt-Kt exp       L t          H t
1     100      31415.93    30798621   30767205.2  3131.055  3031.055025
2     200     125663.71    48583740   48458076.6  3932.516  3732.516350
3     300     282743.34    80249928   79967184.8  5054.141  4754.141421
4     400     502654.82   132737719  132235064.2  6500.133  6100.132940
5     500     785398.16   202143062  201357664.2  8021.480  7521.479612
...
```

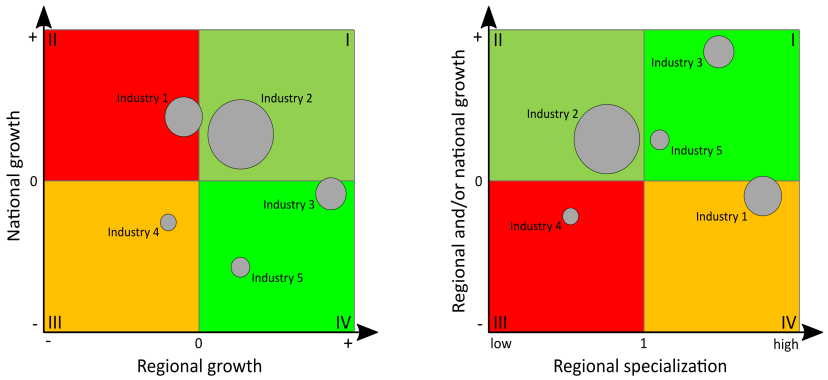
The graphical output is shown in Figures 4a (pharmacies) and 4b (psychotherapists), respectively. The expected value of K_t is plotted as blue line, while the empirical K_t values are red and the corresponding confidence intervals are colored in green (These colors are the default values in `ripley()` and can be changed by the function arguments `lcol.exp` and `lcol.emp`, respectively). As we have nearly the same number of points in both cases within the same field area, a direct comparison seems reasonable. Obviously, both types of locations show a significant spatial clustering: Also the pharmacies are more clustered than expected on condition of complete spatial randomness up to a distance of about 15,000 meters. We have to remember that also the population is already clustered (see Section 2.2.2) and the spatial distribution of pharmacies may follow this pattern. However, the clustering of psychotherapists exceeds this level enormously, especially within smaller distances up to about 8,000 meters. In conclusion, the psychotherapists are more spatially clustered than pharmacies.

6 Analysis and prognosis of regional growth

6.1 Tools and models concerning regional growth

6.1.1 Analyzing regional growth: shift-share analysis and portfolio matrix

Aspects of regional growth have already been discussed in the context of regional convergence in Section 3. The identification of clusters was the topic of Section 4. Combining some aspects of both, this section presents a collection of tools and models concerning regional growth with respect to industries. Like the indicators in Section 4, these techniques are of high significance especially in the context of local economic policy and municipal business promotion activities, aiming at e.g. strengthening a city's or region's competitiveness, defining its profile or increasing the number of jobs (Dinc, 2015; Nischwitz et al., 2017). Inspired by Farhauer, Kröll (2014) and congruent with the mathematical formulations in Section 4, we calculate on the basis of local/regional employment, e_{ij},



(a) Regional growth portfolio

(b) Regional growth & specialization portfolio

Figure 5: Regional economic portfolio matrix

which is the number of employees of industry i in region j. Its growth from time t to time $t + y$ can be operationalized as an absolute value ($\Delta e_{ij} = e_{ij_{t+y}} - e_{ij_t}$) or as a relative growth ($\Delta e_{ij}^{rg} = e_{ij_{t+y}}/e_{ij_t}$) or as a (percentage) growth rate ($\Delta e_{ij}^{gr} = e_{ij_{t+y}}/e_{ij_t} - 1$).

The first technique described is the regional economic portfolio matrix, originating from the portfolio matrix in marketing, developed by the Boston Consulting Group (BCG) for the identification of growing and declining business fields of firms (Henderson, 1973). However, this technique can also be applied to several regional economic contexts (Baker et al., 2002; Howard, 2007). Here, we present a portfolio matrix which compares the growth in *one* region with the growth in a superordinate reference region (e.g. whole economy). When using the matrix in this way, it is a plot of the growth rate with respect to industry i in the region (Δe_{ij}^{gr}) on the x axis and the corresponding growth in the reference region (Δe_{i}^{gr}) on the y axis (see Figure 5a). The size of the points for each industry may be the total size of employment in the region (e_{ij}) to reflect the absolute relevance of the i-th industry. The plot is segmented into four quadrants, differentiated with respect to positive or negative growth rates. As implied by the colors of the quadrants, they can be interpreted as follows: Quadrant I (top right) contains the industries growing in both the region and the whole economy (or any other reference region). Quadrant II (top left) shows all industries growing in the whole economy but shrinking in the regarded region, which may indicate significant locational handicaps. Quadrant III (bottom left) includes all industries shrinking in the region as well as in the whole economy. Quadrant IV (bottom right) shows the special case of "star" industries, indicating that these industries grow in the regarded region while shrinking in the whole economy. Note that this segmentation (and the corresponding interpretation) differs from the original BCG matrix.

Another variant of the portfolio matrix, which was developed in the context of designing the `REAT` package, is shown in Figure 5b. Combining the aspects of regional specialization (see Section 4) and regional growth, we can plot the location quotient as an indicator of local specialization on the x axis, while plotting an industry-specific growth indicator on the y axis. For identifying "growing" industries, there are at least three options of operationalization: We can plot the industry-specific regional growth rate (Δe_{ij}^{gr}) on the y axis (which is on the x axis in the portfolio matrix in Figure 5a) or the industry-specific national rate (Δe_{i}^{gr}) or, if we want to show regional growth in relation to national growth, the quotient of industry-specific regional and national growth rates ($\Delta e_{ij}^{gr}/\Delta e_{i}^{gr}$). In quadrant I, we see now all industries overrepresented in the region (in terms of the location quotient) as well as growing on the regional/national level. Quadrant II shows all industries underrepresented in the region but growing as well. In quadrants III and IV, we can identify all industries with negative growth rates, which are underrepresented or overrepresented, respectively.

Table 12: Shift-share analysis: Dunn and Gerfin type

Component	Dunn-type (absolute)	Gerfin-type (index)
	$\Delta e_j = e_{j_{t+y}} - e_{j_t} = n_{j_{t,t+y}} + m_{j_{t,t+y}} + c_{j_{t,t+y}}$	
Net total shift	$t_{t+y} = e_{j_{t+y}} - e_{j_t} - n_{j_{t,t+y}} = m_{j_{t,t+y}} + c_{j_{t,t+y}}$	$t_{t+y} = m_{j_{t,t+y}} c_{j_{t,t+y}} = \frac{\frac{e_{j_{t+y}}}{e_{j_t}}}{\frac{e_{t+y}}{e_t}}$
static (two time periods t and $t+y$)		
National share	$n_{j_{t,t+y}} = e_{j_t} \frac{e_{t+y}}{e_t} - e_{j_t}$	$n_{j_{t,t+y}} = 1$ (omitted)
Industrial mix	$m_{j_{t,t+y}} = \sum_{i=1}^{I} e_{ij_t} \frac{e_{i_{t+y}}}{e_i} - e_{j_t} \frac{e_{t+y}}{e_t}$	$m_{j_{t,t+y}} = \frac{\sum_{i=1}^{I} e_{ij_t} \frac{e_{i_{t+y}}}{e_i}}{e_{j_t} \frac{e_{t+y}}{e_t}}$
Regional share	$c_{j_{t,t+y}} = \sum_{i=1}^{I} e_{ij_t} (\frac{e_{ij_{t+y}}}{e_{ij_t}} - \frac{e_{i_{t+y}}}{e_{i_t}})$	$c_{j_{t,t+y}} = \frac{\frac{e_{ij_{t+y}}}{e_{ij_t}}}{\sum_{i=1}^{I} e_{ij_t} \frac{e_{i_{t+y}}}{e_i}}$
dynamic (T time periods, while $T > 2$)		
National share	$n_{j_{t,T}} = \sum_{t=1}^{T} e_{j_t} \frac{e_{t+1}}{e_t} - e_{j_t}$	
Industrial mix	$m_{j_{t,T}} = \sum_{t=1}^{T} \sum_{i=1}^{I} e_{ij_t} \frac{e_{i_{t+1}}}{e_i} - e_{j_t} \frac{e_{t+1}}{e_t}$	
Regional share	$c_{j_{t,T}} = \sum_{t=1}^{T} \sum_{i=1}^{I} e_{ij_t} (\frac{e_{ij_{t+1}}}{e_{ij_t}} - \frac{e_{i_{t+1}}}{e_{i_t}})$	
industry-specific		
National share	$n^i_{j_{t,t+y}} = e_{ij_t} \frac{e_{i_{t+y}}}{e_{i_t}} - e_{ij_t}$	
Regional share	$c^i_{j_{t,T}} = e_{ij_t} (\frac{e_{ij_{t+y}}}{e_{ij_t}} - \frac{e_{i_{t+y}}}{e_{i_t}})$	
prognosis for time period z		
Employment		$\Delta e_{ij_{t+z}} = e_{ij_{t+y}} (\frac{e^P_{i_{t+z}}}{e_{i_{t+y}}}) c_{j,t+y}$

Notes: e_{j_t} is the employment in region j at time t, e_{ij_t} is the employment of industry i in region j at time t, e_t is the total employment in the whole economy at time t, e_{i_t} is the total employment in industry i, y and z are numbers of time periods added to t ($z > y$), T is the number of regarded time periods and I is the number of industries.
Compiled from: Farhauer, Kröll (2014); Haynes, Parajuli (2014); Schätzl (2000); Schönebeck (1996); Spiekermann, Wegener (2008); Barff, Knight (1988)

A well-established model of regional growth is the shift-share analysis, which is, although developed independently from the portfolio matrix, closely linked to the concept presented above. The original shift-share analysis was introduced by Dunn Jr. (1960) and given a theoretical foundation by Casler (1989). Parallelly and independently, Gerfin (1964) developed a variant of shift-share analysis, which is more popular in the German-speaking regional economic science. Both concepts have been extended in several ways. Table 12 shows the basics of shift-share analysis with respect to "Dunn" and "Gerfin" type. As there are several ways of formulating the shift-share formulae and calling the particular elements of the shift-share analysis, the description here is based on the mathematical formulations in Farhauer, Kröll (2014) and the terms used in Haynes, Parajuli (2014).

The basic idea of shift-share analysis is the decomposition of regional growth into components, recognizing that single economic regions are embedded into and influenced by a larger regional system, normally the whole economy, just called "the nation" hereinafter: The (employment or e.g. gross value added) growth of industry i in region j from time t to time $t+y$ can be attributed to 1) a national trend, which means the economic climate in the whole system of regions, 2) the all-over growth or decline of the regarded industries and 3) the industry-specific performance of the region, which is linked to locational advantages or disadvantages. The first component is called *national share* and reflects the growth in region j that *would* have occurred if region j *would* have developed exactly as the nation. The second component is the *industrial mix*, represent-

ing the aggregated industry-specific growth in region j if the regarded industries *would* have developed like in the whole economy, adjusted by the national effect. The third component is the *regional share*, which is the "residuum" of the first two components; this share of growth is attributed to locational advantages (or disadvantages), showing the regional growth adjusted by national and industry effects (Farhauer, Kröll, 2014; Haynes, Parajuli, 2014).

The Dunn-type models deal with absolute growth (Δe_{ij} or Δe_j), which is the sum of all shift-share components, and a *net total shift*, which is the sum of the industrial mix and the regional share (as these components are region-specific). Thus, this technique is also called the "difference method". The Gerfin-type approaches express growth in terms of indices, while the net total shift for region j is the result of a multiplication of the industrial mix index and the regional share index, resulting in the alternative denomination "index method" (Schätzl, 2000).

Several extensions have been developed for the Dunn-type shift-share analysis (Haynes, Parajuli, 2014). One regular application calculates a shift-share analysis for each industry i in region j (instead of computing components for the whole region), while skipping the industrial mix effect. A main contribution was the dynamic shift-share analysis by Barff, Knight (1988). It extended the Dunn model by dealing with growth within a longitudinal cut of T years. Other extensions of the Dunn-type technique provide a deeper differentiation of the three components, which are regarded as correlated (e.g. Arcelus 1984; Esteban-Marquillas 1972).

6.1.2 Commercial area prognosis

Also developed independently in the context of German urban planning, a commercial area prognosis deals with an absolute (assumed) employment growth (Δe_{ij}) over T years, which is used to forecast the required commercial area within a city or region j up to time T. Note that "commercial area" represents the type of urban area which is used by specific economic activities, especially industrial plants, and/or designated for this purpose in municipal land-use plans. This technique is a demand-side approach, as it derives the required commercial area from the (expected) demand for it (Bonny, Kahnert, 2005). See Table 13 for the calculation of two types of commercial area prognosis based on employment growth.

The basic model called *GIFPRO* (German abbreviation for "Gewerbe- und Industrieflächenbedarfsprognose", roughly translated: prognosis of future demand of commercial area) was developed by Stark et al. (1981). The usual procedure is to estimate – starting from the current employment – the future industry-specific employment in region j. This number of employees is weighted by the industry-specific shares of workers usually located in commercial areas and multiplied by a resettlement rate (sq_{ij} percent of employees from industry i are resettled in one time period) and a relocation rate (rq_{ij} percent of employees from industry i are relocated in one time period) as well as a reutilization rate (ru_{ij} percent of employees from industry i will be located at reused commercial area). This "commercial area-relevant" employment is weighted with an areal index, a_{ij} (commercial area per employee), to compute the commercial area for industry i in region j for one time period t. The expected commercial area is summed over all I industries (A_{j_t}) and, finally, over all T years and I industries (A_{j_T}) (Bonny, Kahnert, 2005; Planungsgruppe MWM, 2009).

A significant extension was developed in the context of establishing a land-use plan for Dresden: The *TBS-GIFPRO* (German abbreviation for "Trendbasierte und standortspezifische Gewerbe- und Industrieflächenbedarfsprognose", roughly translated: trend-based and location-specific prognosis of future demand of commercial area) technique (Deutsches Institut für Urbanistik GmbH, Spath + Nagel (GbR), 2010). It includes a stochastic approach for forecasting employment as well as other region-specific data. The employment prognosis is done using a trend regression model (employment against time) based on past empirical employment data for region j (mostly from official employment statistics) which are used for forecasting future employment. For each i industry, a single regression model is estimated, where the function type is not pre-defined but chosen e.g. based on the explained variance (R^2) and/or plausibility considerations.

Table 13: Commercial area prognosis

Prognosis	GIFPRO	TBS-GIFPRO
Employment	$e^A_{ij_t} = \left[\left(e_{ij_{t0}} \frac{a_i}{100} \frac{sq_{ij}}{100}\right) + \left(e_{ij_{t0}} \frac{a_i}{100} \frac{rq_{ij}}{100}\right)\right] - \left(e_{ij_{t0}} \frac{ru_{ij}}{100}\right)$	$e^A_{ij_t} = \left[\left(e_{ij_t} \frac{a_i}{100} \frac{sq_{ij}}{100}\right) + \left(e_{ij_t} \frac{a_i}{100} \frac{rq_{ij}}{100}\right)\right] - \left(e_{ij_t} \frac{ru_{ij}}{100}\right)$ where: $e_{ij_t} = f(t) = a + bt$ or $f(t) = at^b$ or $f(t) = ae^{bt}$ or $f(t) = \frac{e^{MAX}_{ij}}{1+e^{-a+bt}}$
Areal index	pre-defined: ai_{ij}	empirical estimation: $ai_{ij} = \frac{A_{ij}}{e_{ij}}$
Commercial area	$A_{ij_t} = e^A_{ij_t} ai_{ij}$ $A_{j_t} = \sum_{i=1}^{I} A_{ij_t}$ $A_{j_T} = \sum_{i=1}^{I} \sum_{t=1}^{T} A_{ij_t}$	

Notes: $e^A_{ij_t}$ is the (expected) number of employees of industry i in region j which is located in commercial areas at time t, $e_{ij_{t0}}$ is the employment of industry i in region j at start time $t0$ (empirical value), e_{ij_t} is the (expected) employment of industry i in region j at time t, a_i is the share of employees in industry i which is located in commercial areas, sq_{ij} is the resettlement rate with respect to industry i in region j in one time period, rq_{ij} is the relocation rate with respect to industry i in region j in one time period, ru_{ij} is the reutilization rate with respect to industry i in region j in one time period, ai_{ij} is the areal index with respect to industry i in region j (commercial area per employee), A_{ij_t} is the (expected) commercial area for industry i in region j at time t, A_{j_t} is the (expected) commercial area in region j at time t and A_{j_T} is the sum of the (expected) commercial area in region j over all T time periods.
Compiled from: Bonny, Kahnert (2005); CIMA Projekt + Entwicklung GmbH et al. (2011); Deutsches Institut für Urbanistik GmbH, Spath + Nagel (GbR) (2010); Planungsgruppe MWM (2009); Mulligan (2006); Vallée et al. (2012)

The function may be linear (which seems unrealistic) or not: Deutsches Institut für Urbanistik GmbH, Spath + Nagel (GbR) (2010) use linear and exponential functions. However, from the growth perspective, also a logistic function may be applied (see Mulligan 2006 for a discussion of logistic growth with respect to population). If possible, the areal index and, maybe, other parameters are also estimated empirically for the specific region j (e.g. via firm-level surveys and/or official statistical data).

6.2 Application in REAT

6.2.1 REAT functions for analyzing and forecasting regional growth

Table 14 shows the functions for the analysis of regional growth as implemented in `REAT`. Table 15 presents the functions related to commercial area prognosis. All of these functions require at least current employment data for each industry in the regarded region j, e_{ij}, which may be a single `numeric vector` or the column of a `data frame` or `matrix`. Another similarity of all mentioned functions is the optional argument of the industry names (or codes). If no industry names are stated by the user (default function argument: `industry.names = NULL`), the industries are numbered consecutively. With respect to the function output, all regional growth functions distinguish between a visible and an invisible output (see e.g. Section 3), where the main results are returned automatically and the details are included in the invisible output (mostly a `list` with several entries of type `matrix`).

The portfolio matrix (growth portfolio and growth-specialization portfolio, respectively) can be plotted using the functions `portfolio()` and `locq.growth()`, respectively. The different techniques of shift-share analysis are distributed over five functions (`shift()`, `shiftd()`, `shifti()`, `shiftid()` and `shiftp()`). The usage of portfolio and shift-share functions is similar: In any case, the user needs industry-specific employment data for the regarded region and the reference region (e.g. whole economy) for at least two time periods (e.g. years).

Table 14: REAT functions for analyzing regional growth

Model	REAT function	Mandatory arguments	Optional arguments	Output
Growth portfolio matrix	portfolio()	vectors of e_{ij_t} and $e_{ij_{t+y}}$ and vectors of e_{i_t} and $e_{i_{t+y}}$ or matrix/data frame with e_{ij_t} and e_{i_t} for T years, point size (e.g. $e_{ij_{t+y}}$)	point size factor, industry names	visible: plot, invisible: growth rates (matrix)
Growth and specialization portfolio matrix	locq.growth()	vectors of e_{ij_t} and $e_{ij_{t+y}}$ and vectors of e_{i_t} and $e_{i_{t+y}}$ or matrix/data frame with e_{ij_t} and e_{i_t} for T years, point size (e.g. $e_{ij_{t+y}}$)	point size factor, industry names	visible: plot, invisible: list with portfolio data (matrix), LQ_{ij} (matrix) and growth rates (matrix)
Shift-share analysis	shift()	vectors of e_{ij_t} and $e_{ij_{t+y}}$, vectors of e_{i_t} and $e_{i_{t+y}}$	shift-share method (default: Dunn), industry names, plot components, plot portfolio	visible: matrix with components, invisible: list with components (matrix), growth (matrix) and shift method (char), optional: plot(s)
dynamic	shiftd()	vectors of $e_{ij_{t0}}$ and $e_{i_{t0}}$, matrix/data frame with e_{ij_t} and e_{i_t} for T years	shift-share method (default: Dunn), industry names, plot components, plot portfolio	visible: matrix with annual components, invisible: list with components (matrix), annual components (matrix), growth (matrix) and shift method (char), optional: plot(s)
industry-specific	shifti()	vectors of e_{ij_t} and $e_{ij_{t+y}}$, vectors of e_{i_t} and $e_{i_{t+y}}$	shift-share method (default: Dunn), industry names, plot components, plot portfolio	visible: matrix with industry components, invisible: list with components (matrix), industry components (matrix), growth (matrix) and shift method (char), optional: plot(s)
industry-specific and dynamic	shiftid()	vectors of $e_{ij_{t0}}$ and $e_{i_{t0}}$, matrix/data frame with e_{ij_t} and e_{i_t} for T years	shift-share method (default: Dunn), industry names, plot components, plot portfolio	visible: matrix with industry components, invisible: list with components (matrix), industry components (matrix), growth (matrix) and shift method (char), optional: plot(s)
prognosis	shiftp()	vectors of e_{ij_t} and $e_{ij_{t+y}}$, vectors of e_{i_t} and $e_{i_{t+y}}$, vector of $e^P_{i_{t+z}}$	industry names, plot	visible: matrix with industry components, invisible: list with industry employment prognosis (matrix), components (matrix), industry components (matrix), growth (matrix) and shift method (char), optional: plots

Source: own compilation.

Table 15: REAT functions for commercial area prognosis

Model	REAT function	Mandatory arguments	Optional arguments	Output
GIFPRO	gifpro()	vectors of e_{ij}, a_i, sq_{ij}, rq_{ij} and ai_{ij}, time interval, time base	vector of ru_{ij}, industry names, type of output	visible: total commercial area and (optional) annual values, invisible: list with components (matrices), annual and all-over results (list with two matrices)
TBS-GIFPRO	gifpro.tbs()	vectors of e_{ij_t} for T years, a_i, sq_{ij}, rq_{ij} and ai_{ij}, time interval, time base, trend function types	vector of ru_{ij}, industry names, type of output, employment forecast only	visible: total commercial area and (optional) annual values, invisible: list with components (matrices), annual and all-over results (list with two matrices), industry-specific forecast model results (list with I matrices)

Source: own compilation.

All functions for shift-share analysis (except for shift-share prognosis with `shiftp()`) provide three variants of calculation of the components: The classical Dunn method (default function argument `shift.method="Dunn"`), the Dunn extension by Esteban-Marquillas (1972) (`shift.method="Esteban"`) producing four components instead of three, and the Gerfin method (`shift.method="Gerfin"`). When calculating a dynamic shift-share analysis, the user must choose the function `shiftd()`. Industry-specific components are returned by the function `shifti()`. With `shiftid()` one can combine both approaches. Here, it is important to recognize that the function structure allows a combination of e.g. industry-specific and dynamic components while calculating the components from the Esteban-Marquillas extension of shift-share analysis. Additionally, the shift-share functions may plot a portfolio matrix (function argument `plot.portfolio = TRUE`), allowing portfolio and shift-share analysis at once.

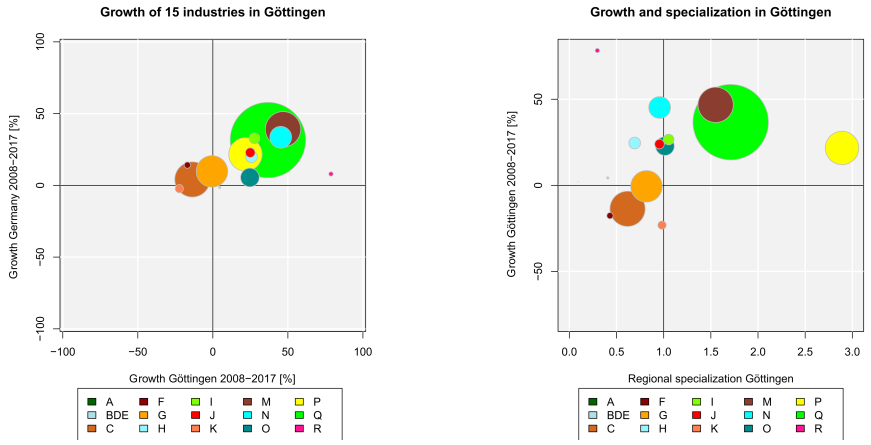
Both functions for commercial area prognosis (`gifpro()` and `gifpro.tbs()`) require vectors of employment data as well as the coefficients for resettlement etc. When forecasting commercial area using the trend-specific technique with `gifpro.tbs()`, the user needs time series data of previous industry-specific employment and has to specify a trend function type (linear, power, exponential or logistic) for each industry. The "best" function type may be examined visually by regarding the employment forecasting output (optional function argument `prog.plot = TRUE`) and the related R^2 values which is part of the invisible function output. Note that this function uses the `REAT` function `curvefit()`, which is a simple tool for bivariate regression, similar to the curve fitting functions in other spreadsheet or statistics software.

6.2.2 Application example 1: Analysis of regional growth in Göttingen

Referring to the example in Section 4.2.2, we perform a regional growth analysis for the German city Göttingen. We use the same dataset `Goettingen` as before, that contains industry-specific employment data for Göttingen and Germany from 2008 to 2017. We load our example data:

```
data(Goettingen)
```

In the first step, we want to examine the industry-specific growth in Göttingen visually. Using the function `portfolio()`, we plot a regional growth matrix with respect to the 15 industries (rows 2 to 16). We also set a plot title (argument `pmtitle`) and axis labels (arguments `pmx` and `pmy`, respectively) as well as industry-specific colors (argument `pcol`):



(a) Growth portfolio matrix for Göttingen (b) Growth-specialization matrix for Göttingen

Figure 6: Portfolio matrix analysis for 15 industries in Göttingen

```
portfolio (Goettingen$Goettingen2008[2:16],
Goettingen$Goettingen2017[2:16],
Goettingen$BRD2008[2:16], Goettingen$BRD2017[2:16],
psize = Goettingen$Goettingen2017[2:16], psize.factor = 15,
pmtitle = "Growth of 15 industries in Göttingen",
industry.names = Goettingen$WZ2008_Code[2:16],
pmx = "Growth Göttingen 2008-2017 [%]",
pmy = "Growth Germany 2008-2017 [%]",
pcol.border = "grey",
pcol = c("darkgreen", "powderblue", "chocolate", "darkred",
"orange", "cadetblue1", "chartreuse1", "red",  "coral",
"coral4", "cyan", "darkcyan", "yellow", "green", "deeppink"),
leg = TRUE, leg.x = -90)
```

Similarly, we plot a growth-specialization portfolio matrix using `locq.growth()` with the same options (colors etc.). On the y axis, we put the industry-specific regional growth which is stated by the function argument `y.axis = "r"` (if we would like to see the national growth instead, we had to set `y.axis = "n"`; for the quotient of regional and national growth, use `y.axis = "rn"`):

```
locq.growth (Goettingen$Goettingen2008[2:16],
Goettingen$Goettingen2017[2:16],
Goettingen$BRD2008[2:16], Goettingen$BRD2017[2:16],
psize = Goettingen$Goettingen2017[2:16], psize.factor = 15,
y.axis = "r", industry.names = Goettingen$WZ2008_Code[2:16],
pmtitle = "Growth and specialization in Göttingen",
pmx = "Regional specialization Göttingen",
pmy = "Growth Göttingen 2008-2017 [%]", pcol.border = "grey",
pcol = c("darkgreen", "powderblue", "chocolate", "darkred",
"orange", "cadetblue1", "chartreuse1", "red", "coral",
"coral4", "cyan", "darkcyan", "yellow", "green", "deeppink"),
leg = TRUE, leg.x = 0.1)
```

The resulting growth portfolio matrix is shown in Figure 6a, the growth-specialization portfolio in Figure 6b. The size of the points (or bubbles) is equal to the current industry-specific employment (e_{ij}) for 2017 (rows 2 to 16 of column `Goettingen2017` in the example data), normalized with respect to a maximum point size of 15 (argument `psize.factor = 15`). As we can see, the health sector (industry code Q, green bubble) has the highest absolute relevance, which can be attributed to the local university hospital (see Section 4.2.2). The axes in the growth portfolio are segmented at $x = 0$ and $y = 0$, respectively, which means a differentiation between positive and negative growth.

As we can see, most industries have grown from 2008 to 2017 in both the region and the whole economy (see quadrant I) with similar growth rates. There is one outlier: Industry R (arts, entertainment, and recreation) shows a regional growth of more than 75 percent, while the national growth is about 10 percent. Note that we see percentage growth rates from 2008 to 2017 here (if *average* growth rates are desired, use the function argument `time.periods`).

Looking at the growth-specialization portfolio, we can identify absolute relevance and growth rate as well as regional specialization of the industries (The colors and bubble sizes are equal to those in Figure 6a). In quadrant I, we find the industries which are overrepresented in Göttingen (specialization) and growing at this regional level. As expected in this university city and related to our results in Section 4.2.2, the "stars" in Göttingen are education (code P), health (code Q) and professional, scientific and technical services (code M).

While the portfolio matrix analysis tells us about the industry-specific growth, the shift-share analysis decomposes this growth into the national, industrial and regional components. In the first step, we perform a static shift-share analysis in the sense of Dunn Jr. (1960) for the same data as in the portfolio analysis by applying the function `shift()`:

```
shift(Goettingen$Goettingen2008[2:16], Goettingen$Goettingen2017[2:16],
Goettingen$BRD2008[2:16], Goettingen$BRD2017[2:16])
# rows 2-6: 15 industries
# columns Goettingen2008 and Goettingen2017:
# employment Goettingen 2008 and 2017, respectively
# columns BRD2008 and BRD2017:
# employment Germany 2008 and 2017, respectively
```

This is our (visible) output:

```
Shift-Share Analysis
Method: Dunn

Shift-share components
                Components
Growth (t1-t)   10411.0000
National share   9178.1916
Industrial mix   2204.8202
Regional share   -972.0118
Net total shift  1232.8084

Calculation for 15 industries
Regional employment at time t: 56872, at time t+1:
67283 (10411 / 18.30602 %)
National employment at time t: 27695398, at time t+1:
32164973 (4469575 / 16.13833 %)
```

In this cross-sectional analysis, we see that the overall employment in Göttingen increased by 10,411 persons from 2008 to 2017. However, a large share of this growth is due to the growth in the national economy ($n_{j_{t,t+y}} \approx 9,178$ employees), which is only a bit lower than Göttingen. The industrial mix component ($m_{j_{t,t+y}}$) shows that approximately 2,205 additional employees must be attributed to an overrepresentation of growing industries in Göttingen. The regional share is negative ($c_{j_{t,t+y}} \approx -972$), which indicates locational disadvantages. When interpreting the industrial mix also as a regional aspect (which seems plausible), we can look at the sum of the industrial mix and the regional share: The net total shift (t_{t+y}) is equal to 1,233 employees, representing the growth difference between the region and the whole economy.

We confirm our results using the Gerfin technique. We request it by setting the argument `shift.method` of the `shift()` function equal to `"Gerfin"`:

```
shift(Goettingen$Goettingen2008[2:16], Goettingen$Goettingen2017[2:16],
Goettingen$BRD2008[2:16], Goettingen$BRD2017[2:16],
shift.method = "Gerfin")
```

The output is:

```
Shift-Share Analysis
Method: Gerfin

Shift-share components
               Components
Industrial mix   1.0333810
Regional share   0.9857591
Net total shift  1.0186647

Calculation for 15 industries
Regional employment at time t: 56872, at time t+1:
67283 (10411 / 18.30602 %)
National employment at time t: 27695398, at time t+1:
32164973 (4469575 / 16.13833 %)
```

In the index method, there is no national share component (implicitly, it is equal to one), thus, we only take a look at the industrial mix and the regional share as well as the net total shift. The industrial mix component is above one ($n_{j_{t,t+y}} \approx 1.03$), showing a more advantageous sector structure in Göttingen compared to Germany. While the regional share in the Dunn-type shift-share analysis was negative, this component in the Gerfin analysis is slightly below one ($c_{j_{t,t+y}} \approx 0.99$), indicating locational disadvantages.

These traditional techniques only regard the overall growth with respect to cross-sectional data. To gain a deeper insight and take into account also seasonal effects, we perform a dynamic shift-share analysis in the sense of Barff, Knight (1988) which distinguishes between the 15 industries simultaneously. This can be done via the `REAT` function `shiftid()`, requiring data for the initial time period and at least for two following periods. In the `Goettingen` dataset, the rows 2 to 16 represent the industries and the columns represent the years (2008 to 2017). Data for the regarded region and the whole economy is arranged successively. We also use the industry codes in column `WZ2008_Code`. In this function, we have to define the start and end periods explicitly:

```
shiftid(Goettingen$Goettingen2008[2:16], Goettingen[2:16,3:12],
Goettingen$BRD2008[2:16], Goettingen[2:16,13:22],
time1 = 2008, time2 = 2017,
industry.names = Goettingen$WZ2008_Code[2:16])
# columns 3-12: employment in Göttingen 2009-2017
# columns 13-22: employment in Germany 2009-2017
```

The result is:

```
Dynamic Shift-Share Analysis
Method: Dunn

Shift-share components
                         A      BDE          C         F         G
Growth (t1-t)   -3.000000 29.00000 -1117.0000 -255.0000  -51.0000
National share   6.103502 -9.46377   254.5217  160.0638  561.7436
Regional share  -9.103502 38.46377 -1371.5217 -415.0638 -612.7436
Net total shift -9.103502 38.46377 -1371.5217 -415.0638 -612.7436
                        H         I         J           K        M
Growth (t1-t)   524.0000 470.00000 274.00000 -465.000000 2229.000
National share  368.2053 515.03493 286.32383    6.356612 1821.392
Regional share  155.7947 -45.03493 -12.32383 -471.356612  407.608
Net total shift 155.7947 -45.03493 -12.32383 -471.356612  407.608
                        N        O         P        Q         R
Growth (t1-t)   1178.0000 268.0000 1272.0000 4211.000 363.00000
National share   977.9869 167.9118 1138.5383 3556.692  47.50353
Regional share   200.0131 100.0882  133.4617  654.308 315.49647
Net total shift  200.0131 100.0882  133.4617  654.308 315.49647
```

```
Calculation for 15 industries
Regional employment at time t: 56872, at time t+1:
67283 (10411 / 18.30602 %)
National employment at time t: 27695398, at time t+1:
32164973 (4469575 / 16.13833 %)
```

The visible output is a `matrix` containing one row for each component (the number of components depends on the selected shift-share method, here: Dunn) and I columns (one for each industry). As we calculate industry-specific components, there is no industrial mix effect, which means that the calculations are on the level of single industries. Again, we detect large absolute growth for industries P (education) and Q (health) (see Table 9). Interestingly, this growth can be mainly attributed to effects in the whole economy. The corresponding regional shares are small but positive, showing locational advantages with respect to these industries in Göttingen.

The logic of shift-share analysis can also be regarded in two other examples: If industry C (manufacturing) *had* developed as in the national trend, the absolute growth in Göttingen *would* be equal to 255 employees. In fact, there was a decline of 1,117 employees, resulting in a negative regional share of -1,372 employees, indicating locational disadvantages with respect to the manufacturing sector. The opposite is true for the industries with code BDE (including electricity, gas, water supply, etc.): The absolute growth of 29 employees *would* not have occurred if this sector *had* developed as in the whole economy (negative national share equal to -9 employees). The residuum (regional share) is equal to 38 employees, indicating a trend contrary to the national.

6.2.3 Application example 2: Commercial area prognosis for Göttingen

Using the same data, we now perform a commercial area prognosis for Göttingen. We load our data:

```
data(Goettingen)
```

When using the GIFPRO-based commercial area prognosis techniques, several parameters have to be defined (employment shares in commercial areas a_i, resettlement rate sq_{ij}, relocation rate rq_{ij} and areal index ai_{ij}; a reutilization rate ru_{ij} is optional, thus, we ignore the reutilization of commercial area in this example). These parameters have to be defined for each industry. In our example, we use the employment shares as well as the resettlement and relocation rates from Deutsches Institut für Urbanistik GmbH, Spath + Nagel (GbR) (2010). Note that some sectors are, per definition, not located within commercial areas (e.g. agriculture), resulting in an employment share of $a_i = 0$. As we want to reuse the sets of parameters, we save them as single `numeric vectors`:

```
ca_share <- c(0, 0, 100, 90, 70, 100, 10, 20, 20, 20, 20, 0, 0, 0, 0)
# industry-specific shares of employees in commercial areas
sq_quote <- c(0.77, 0.77, 0.15, 0.15, 0.77, 0.15, 0.77, 0.77,
0.77, 0.77, 0.77, 0.77, 0.77, 0.77, 0.77)
# industry-specific resettlement quote
rq_quote <- rep(0.7, 15)
# industry-specific relocation quote (0.7 for each of the 15 industries)
area_index <- c(0, 0, 200, 75, 250, 250, 50, 100, 100, 100, 100,
50, 50, 50, 50)
# industry-specific area index (sqm commercial area per employee)
```

Now, we compute the traditional commercial area prognosis using the `gifpro()` function and the `Goettingen` data as well as the parameters defined above. We forecast the commercial area for five years (`tinterval = 5`). Our base is 2017 (`time.base = 2017`), as this is the last year empirical data is available for. We save the (invisible) output in the `list` object `gifpro_goettingen`:

```
gifpro_goettingen <- gifpro (e_ij = Goettingen$Goettingen2017[2:16],
a_i = ca_share, sq_ij = sq_quote, rq_ij = rq_quote, tinterval = 5,
ai_ij = area_index, time.base = 2017,
industry.names = Goettingen$WZ2008_Code[2:16], output = "full")
```

As we have set `output = "full"`, the visible function output contains overall as well as annual values:

```
GIFPRO
Method: GIFPRO

Employment and commercial area changes (allover)
        Employment Commercial Area
Sum      1113.8785       212981.94
Average   222.7757        42596.39

Employment and commercial area changes (per time unit)
     Employment CommercialArea
2018   222.7757       42596.39
2019   222.7757       42596.39
2020   222.7757       42596.39
2021   222.7757       42596.39
2022   222.7757       42596.39

Calculation for 15 industries
```

In all 15 industries, 1,114 new employees are predicted for the year 2022, resulting in 212,928 square meters required for new commercial area. As the employment prognosis is not based on (nonlinear) trend regression but on constant growth, the absolute employment growth and the required commercial area are equal in each year (223 employees and 42,596 sqm, respectively).

The object `gifpro_goettingen` contains a `list` called `components` containing the single components of prognosis as well as the results already shown in the visible output (`results`). To understand the GIFPRO technique and the related `REAT` function, we take a look at the single components:

```
gifpro_goettingen$components

$resettlement
        2018     2019     2020     2021     2022
A    0.00000  0.00000  0.00000  0.00000  0.00000
BDE  0.00000  0.00000  0.00000  0.00000  0.00000
C   11.81100 11.81100 11.81100 11.81100 11.81100
F    1.80090  1.80090  1.80090  1.80090  1.80090
G   38.00489 38.00489 38.00489 38.00489 38.00489
H    3.72150  3.72150  3.72150  3.72150  3.72150
I    1.73327  1.73327  1.73327  1.73327  1.73327
J    3.12928  3.12928  3.12928  3.12928  3.12928
K    2.67806  2.67806  2.67806  2.67806  2.67806
M   12.18910 12.18910 12.18910 12.18910 12.18910
N    7.50750  7.50750  7.50750  7.50750  7.50750
O    0.00000  0.00000  0.00000  0.00000  0.00000
P    0.00000  0.00000  0.00000  0.00000  0.00000
Q    0.00000  0.00000  0.00000  0.00000  0.00000
R    0.00000  0.00000  0.00000  0.00000  0.00000

$relocation
       2018    2019    2020    2021    2022
A    0.0000  0.0000  0.0000  0.0000  0.0000
BDE  0.0000  0.0000  0.0000  0.0000  0.0000
C   55.1180 55.1180 55.1180 55.1180 55.1180
F    8.4042  8.4042  8.4042  8.4042  8.4042
G   34.5499 34.5499 34.5499 34.5499 34.5499
H   17.3670 17.3670 17.3670 17.3670 17.3670
I    1.5757  1.5757  1.5757  1.5757  1.5757
J    2.8448  2.8448  2.8448  2.8448  2.8448
K    2.4346  2.4346  2.4346  2.4346  2.4346
```

```
M   11.0810 11.0810 11.0810 11.0810 11.0810
N    6.8250  6.8250  6.8250  6.8250  6.8250
O    0.0000  0.0000  0.0000  0.0000  0.0000
P    0.0000  0.0000  0.0000  0.0000  0.0000
Q    0.0000  0.0000  0.0000  0.0000  0.0000
R    0.0000  0.0000  0.0000  0.0000  0.0000

$reuse
    2018 2019 2020 2021 2022
A      0    0    0    0    0
BDE    0    0    0    0    0
C      0    0    0    0    0
F      0    0    0    0    0
G      0    0    0    0    0
H      0    0    0    0    0
I      0    0    0    0    0
J      0    0    0    0    0
K      0    0    0    0    0
M      0    0    0    0    0
N      0    0    0    0    0
O      0    0    0    0    0
P      0    0    0    0    0
Q      0    0    0    0    0
R      0    0    0    0    0

$employment
        2018     2019     2020     2021     2022
A    0.00000  0.00000  0.00000  0.00000  0.00000
BDE  0.00000  0.00000  0.00000  0.00000  0.00000
C   66.92900 66.92900 66.92900 66.92900 66.92900
F   10.20510 10.20510 10.20510 10.20510 10.20510
G   72.55479 72.55479 72.55479 72.55479 72.55479
H   21.08850 21.08850 21.08850 21.08850 21.08850
I    3.30897  3.30897  3.30897  3.30897  3.30897
J    5.97408  5.97408  5.97408  5.97408  5.97408
K    5.11266  5.11266  5.11266  5.11266  5.11266
M   23.27010 23.27010 23.27010 23.27010 23.27010
N   14.33250 14.33250 14.33250 14.33250 14.33250
O    0.00000  0.00000  0.00000  0.00000  0.00000
P    0.00000  0.00000  0.00000  0.00000  0.00000
Q    0.00000  0.00000  0.00000  0.00000  0.00000
R    0.00000  0.00000  0.00000  0.00000  0.00000
```

As we defined some industries as not relevant for commercial areas ($a_i = 0$), they do not contribute any employees neither resettled nor relocated (such as A - agriculture, B - mining and quarrying or R - arts, entertainment, and recreation). We see that e.g. in the manufacturing sector (code C), there is an annual increase of about 12 employees attributed to resettlement and 55 employees related to relocation each year (see row 3 in `resettlement` and `relocation`, respectively). As we ignored the reutilization of commercial area, the `matrix` containing the commercial area-relevant employment related to reutilization (`reuse`) contains only zeros. The sum of all three components is stored in the fourth `matrix`, `employment`. There is an annual increase of nearly 67 employees in the manufacturing sector. The contents of the `results list` is the same as shown in the visible output.

In the next step, we apply the trend-based commercial area prognosis (TBS-GIFPRO) to the `Goettingen` data. In the `gifpro.tbs()` function, we use the employment data from 2008 to 2017 (columns 3 to 12), and assume an exponential function for employment prognosis (function argument `prog.func`, repeating the argument `"exp"` for each industry). The employment prognosis is plotted (`prog.plot = TRUE`), showing all 15 plots in one (`plot.single = FALSE`):

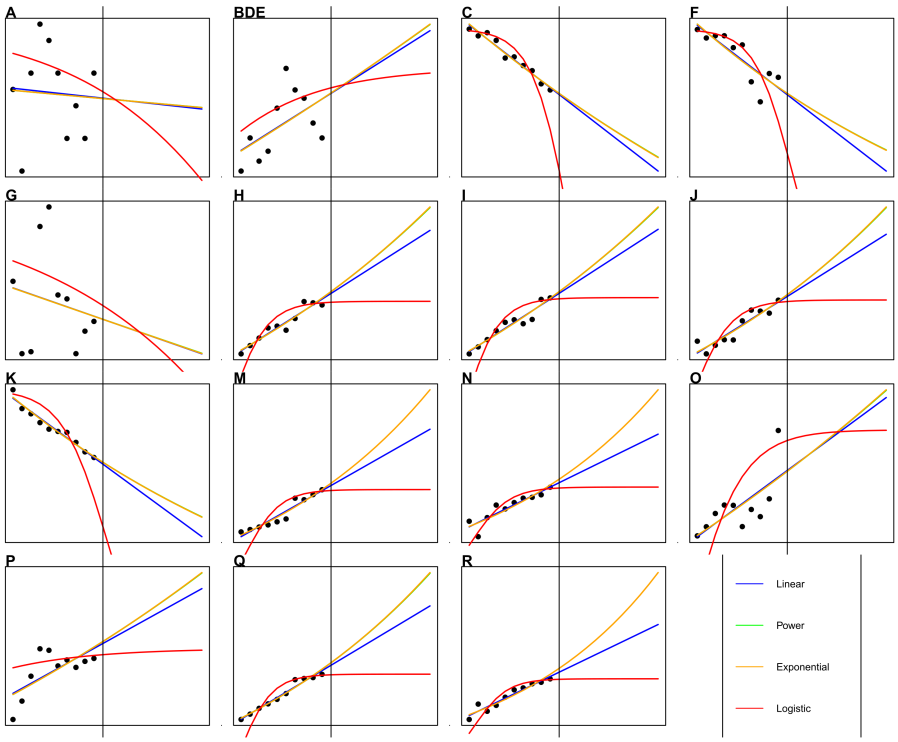


Figure 7: Employment prognosis for 15 industries in Goettingen (TBS-GIFPRO)

```
gifpro.tbs (e_ij = Goettingen[2:16,3:12],
a_i = ca_share, sq_ij = sq_quote, rq_ij = rq_quote, tinterval = 5,
prog.func = rep("exp", nrow(Goettingen[2:16,3:12])),
ai_ij = area_index, time.base = 2008,
industry.names = Goettingen$WZ2008_Code[2:16],
prog.plot = TRUE, plot.single = FALSE, output = "full")
```

The visible function output is similar to the output above:

```
GIFPRO
Method: TBS-GIFPRO

Employment and commercial area changes (allover)
        Employment CommercialArea
Sum      1139.6592      216012.46
Average   227.9318       43202.49

Employment and commercial area changes (per time unit)
     Employment CommercialArea
2018   224.9565       42945.53
2019   226.2904       43054.76
2020   227.7755       43182.97
2021   229.4169       43330.70
2022   231.2199       43498.50

Calculation for 10 industries
```

The resulting plot containing the employment forecasting functions is shown in Figure 7. The black vertical lines divide the plots into the esimation segment (2008 to 2017) and the prognosis segment (2018 to 2022). Four function types are supplied: linear (blue), power (green), exponential (yellow) and logistic (red). Note that a linear trend seems unrealistic as it implies continuous growth and may result in negative employment if the slope is negative. At this point, we should normally discuss and find the "best"

forecasting model for each industry and rerun our analysis a few times. In our example, we skip this step and just take a look at the prognosis functions: In most cases, an exponential growth (or decline) seems to be an appropriate approximation. The power functions (green lines) are nearly invisible as their data fit is nearly the same as that of the exponential functions. Thus, we could choose them instead. In our case, the exponential function seems sufficient.

As expected, a nonlinear industry growth results in a nonlinear overall employment growth and, consequently, the commercial area-relevant employment also grows in a nonlinear way. As we can see from the `gifpro.tbs()` output, employment increases by about 228 employees per year on average and by about 1,140 employees over the five years regarded (2018 to 2022). The annual commercial area required ranges from 42,946 sqm (2018) to 43,499 sqm (2022), all in all 216,012 sqm up to 2022. In our case, the estimated commercial area exceeds the prognosis derived from the simple GIFPRO analysis, which can be attributed to the positive differences between the exponential prognosis and a linear prognosis (see Figure 7). We skip the inspections of the components, which could be addressed by saving the results in an object (`list`), as we did in the first GIFPRO example.

7 Final remarks

This paper has shown how `R` and specifically the package `REAT` can be used for regional economic analysis. It should be noted that this package aims at width with respect to the treated analysis subjects rather than depth. The subsections provide the basic analysis methods regarded as most important from the package developer's point of view (with respect to usage in current papers and discussion in current textbooks as well as application in own research projects), while there are several other approaches as well as extensions of the basic methods. A more detailed survey of the common methods can be found in the cited literature, especially in review articles (e.g. Nakamura, Morrison Paul 2009; Portnov, Felsenstein 2010) and textbooks (e.g. Farhauer, Kröll 2014).

Finally, we have to keep in mind that this package (like nearly any other free software) was developed in a non-commercial context (and published under the GNU General Public License). All functions have been tested several times using various real data and single functions have already been used in a few research projects. However, there is no warranty that all functions always work perfectly. Like nearly any other `R` package, `REAT` is continuously refined, which means extending functions as well as correcting errors. This requires attentive usage and, of course, constructive feedback from the package users. It can be easily transmitted using the contact information on the CRAN package website.

References

Albacete X, Olaru D, Paül V, Biermann S (2017) Measuring the accessibility of public transport: A critical comparison between methods in Helsinki. *Applied Spatial Analysis and Policy* 10[2]: 161–188. CrossRef.

Allington NF, McCombie J (2007) Economic Growth and Beta-Convergence in the East European Transition Economies. In: Arestis P, Baddeley M, McCombie J (eds), *Economic Growth*. Edward Elgar publishing, Cheltenham, 200–222

Arcelus FJ (1984) An extension of shift-share analysis. *Growth and Change* 15[1]: 3–8. CrossRef.

Bai CE, Tao Z, Tong YS (2008) Bureaucratic integration and regional specialization in China. *China Economic Review* 19[2]: 308–319. CrossRef.

Baker P, von Kirchbach F, Mimouni M, Pasteels JM (2002) Analytical Tools for Enhancing the Participation of Developing Countries in the Multilateral Trading System in the Context of the Doha Development Agenda. *Aussenwirtschaft* 57[3]: 343–372. https://EconPapers.repec.org/RePEc:usg:auswrt:2002:57:03:343-372

Balassa B (1965) Trade Liberalisation and "Revealed" Comparative Advantage. *The Manchester School* 33[2]: 99–123. CrossRef.

Barff RA, Knight PL (1988) Dynamic shift-share analysis. *Growth and Change* 19[2]: 1–10. CrossRef.

Barro RJ, Sala-i Martin X (2004) *Economic Growth* (2nd ed.). MIT Press

Bonny HW, Kahnert R (2005) Zur Ermittlung des Gewerbeflächenbedarfs. *Raumforschung und Raumordnung* 63[3]: 232–240

Capello R, Nijkamp P (2009) Introduction: Regional growth and development theories in the twenty-first century - recent theoretical advances and future challenges. In: Capello R, Nijkamp P (eds), *Handbook of Regional Growth and Development Theories*. 1–18

Casler SD (1989) A Theoretical Context for Shift and Share Analysis. *Regional Studies* 23[1]: 43–48. CrossRef.

Ceapraz IL (2008) The Concepts of Specialisation and Spatial Concentration and the Process of Economic Integration: Theoretical Relevance and Statistical Measures. The Case of Romania's Regions. *Romanian Journal of Regional Science* 2[1]: 68–93

Charles-Coll JA (2011) Unterstanding Income Equality: Concept, Causes and Management. *International Journal of Economics and Management Science* 1[3]: 17–28

CIMA Projekt + Entwicklung GmbH, NIW Niedersächsisches Institut für Wirtschaftsforschung, NORD/LB Regionalwirtschaft, Planquadrat Dortmund GbR (2011) Gewerbeflächenkonzeption für die Metropolregion Hamburg (GEFEK). Research report

Cracau D, Durán Lima JE (2016) On the Normalized Herfindahl-Hirschman Index: A Technical Note. *International Journal on Food System Dynamics* 7[4]: 382–386

Damgaard C, Weiner J (2000) Describing inequality in plant size or fecundity. *Ecology* 81[4]: 1139–1142. CrossRef.

Dapena AD, Fernández Vázquez E, Rubiera Morollón F (2016) The role of spatial scale in regional convergence: the effect of MAUP in the estimation of β-convergence equations. *The Annals of Regional Science* 56[2]: 473–489. CrossRef.

Dauth W, Fuchs M, Otto A (2015) Standortmuster in Westdeutschland: Nur wenige Branchen sind räumlich stark konzentriert. IAB Kurzbericht 16/2015, Institut für Arbeitsmarkt- und Berufsforschung. http://doku.iab.de/kurzber/2015/kb1615.pdf

Dauth W, Fuchs M, Otto A (2018) Long-run processes of geographical concentration and dispersion: Evidence from Germany. *Papers in Regional Science* 97[3]: 569–593. CrossRef.

Deutsches Institut für Urbanistik GmbH, Spath + Nagel (GbR) (2010) Stadtentwicklungskonzept Gewerbe für die Landeshauptstadt Potsdam. Research report, Landeshauptstadt Potsdam. https://www.potsdam.de/sites/default/files/documents/STEK_Gewerbe_Langfassung_2010.pdf

Dinc M (2015) *Introduction to Regional Economic Development. Major Theories and Basic Analytical Tools*. Elgar

Dixon R, Freebairn J (2009) Trends in Regional Specialisation in Australia. *Australasian Journal of Regional Studies* 15[3]: 281–296

Doran J, Jordan D (2013) Decomposing European NUTS2 regional inequality from 1980 to 2009: National and European policy implications. *Journal of Economic Studies* 40[1]: 22–38. CrossRef.

Dunn Jr. ES (1960) A Statistical and Analytical Technique for Regional Analysis. *Papers in Regional Science* 6[1]: 97–112. CrossRef.

Duranton G, Puga D (2000) Diversity and Specialisation in Cities: Why, Where and When Does it Matter? *Urban Studies* 37[3]: 533–555. CrossRef.

Ellison G, Glaeser E (1997) Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy* 105[5]: 889–927

Espa G, Arbia G, Giuliani D (2010) Measuring industrial agglomeration with inhomogeneous K-function: the case of ICT firms in Milan (Italy). Department of Economics Working Papers 1014, Department of Economics, University of Trento, Italia

Esteban-Marquillas JM (1972) I. A reinterpretation of shift-share analysis. *Regional and Urban Economics* 2[3]: 249 – 255. CrossRef.

Farhauer O, Kröll A (2014) *Standorttheorien. Regional- und Standortökonomik in Theorie und Praxis* (2nd ed.). Springer, Heidelberg

Fujita M, Krugman P, Venables A (2001) *The Spatial Economy: Cities, Regions, and International Trade* (1st ed.), Volume 1. The MIT Press

Fülöp G, Kopetsch T, Schöpe P (2011) Catchment areas of medical practices and the role played by geographical distance in the patient's choice of doctor. *The Annals of Regional Science* 46[3]: 691–706. CrossRef.

Furceri D (2005) Beta and sigma convergence: A mathematical relation of causality. *Economics Letters* 89[2]: 212–215. CrossRef.

Gerfin H (1964) Gesamtwirtschaftliches Wachstum und regionale Entwicklung. *Kyklos* 17[4]: 565–593. CrossRef.

Gini C (1912) *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. Cuppini

Gluschenko K (2018) Measuring regional inequality: to weight or not to weight? *Spatial Economic Analysis* 13[1]: 36–59. CrossRef.

Goecke H, Hüther M (2016) Regional Convergence in Europe. *Intereconomics* 51[3]: 165–171. CrossRef.

Goschin Z, Constantin D, Roman M, Ileanu B (2009) Regional specialization and geographic concentration of industries in Romania. *South-Eastern Europe Journal of Economics* 1[1]: 99–113. https://ojs.lib.uom.gr/index.php/seeje/article/view/5536

Haas A, Südekum J (2005) Spezialisierung und Branchenkonzentration in Deutschland: Regionalanalyse. IAB-Kurzbericht 1/2005. http://hdl.handle.net/10419/158181

Habánik J, Hošták P, Kútik J (2013) Economic and social disparity development within regional development of the Slovak Republic. *Economics and Management* 18[3]: 457–464. CrossRef.

Hansen WG (1959) How Accessibility Shapes Land Use. *Journal of the American Institute of Planners* 25[2]: 73–76. CrossRef.

Harris CD (1954) The Market as a Factor in the Localization of Industry in the United States. *Annals of the Association of American Geographers* 44[4]: 315–348

Haynes KE, Parajuli J (2014) Shift-share analysis: decomposition of spatially integrated systems. In: *Handbook of Research Methods and Applications in Spatially Integrated Social Science*. Elgar, 315–344. CrossRef.

Heinemann M (2008) Messung und Darstellung von Ungleichheit. Working Paper Series in Economics 108, University of Lüneburg, Institute of Economics. https://EconPapers.repec.org/RePEc:lue:wpaper:108

Henderson BD (1973) The Experience Curve - Reviewed. IV. The Growth Share Matrix or The Product Portfolio. Reprint 135. https://www.bcg.com/documents/file13904.pdf

Herfindahl OC (1950) *Concentration in the U.S. Steel Industry*. Colombia University Press

Hirschman AO (1945) *National Power and the Structure of Foreign Trade*. Publications of the Bureau of Business and Economic Research. University of California Press

Hoen AR, Oosterhaven J (2006) On the measure of comparative advantage. *The Annals of Regional Science* 40[3]: 677–691. CrossRef.

Hoffmann J, Hirsch S, Simons J (2017) Identification of spatial agglomerations in the German food processing industry. *Papers in Regional Science* 96[1]: 139–162. CrossRef.

Hoover EM (1936) The Measurement of Industrial Localization. *The Review of Economics and Statistics* 18[4]: 162–171

Howard D (2007) A regional economic performance matrix – an aid to regional economic policy development. *Journal of Economic and Social Policy* 11[2]: article 4. https://EconPapers.repec.org/RePEc:usg:auswrt:2002:57:03:343-372

Howard E, Newman C, Tarp F (2016) Measuring industry coagglomeration and identifying the driving forces. *Journal of Economic Geography* 16[5]: 1055–1078

Huang Y, Leung Y (2009) Measuring Regional Inequality: A Comparison of Coefficient of Variation and Hoover Concentration Index. *The Open Geography Journal* 2[1]: 25–34. CrossRef.

Jiang L, Guan M, Tian J (2007) On Chinese Regional Specialization and Industry Concentration. In: *2007 International Conference on Machine Learning and Cybernetics*, Volume 6, 3396–3400

Kabacoff RI (2017) Quick-R: Data Types. Manual. https://www.statmethods.net/input/datatypes.html

Kassenärztliche Bundesvereinigung (2013) Die neue Bedarfsplanung. Grundlagen, Instrumente und regionale Möglichkeiten. Brochure. https://www.kbv.de/media/sp/Instrumente_Bedarfsplanung_Broschuere.pdf

Kim S (1995) Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in U. S. Regional Manufacturing Structure, 1860–1987. *The Quarterly Journal of Economics* 110[4]: 881–908. CrossRef.

Kiskowski MA, Hancock JF, Kenworthy A (2009) On the Use of Ripley's K-function and its Derivatives to Analyze Domain Skriderize. *Biophysical Journal* 97[4]: 1095–1103. CrossRef.

Kohn W, Öztürk R (2013) *Statistik für Ökonomen. Datenanalyse mit R und SPSS* (2nd ed.). Springer Gabler

Krider R, Putler DS (2013) Which Birds of a Feather Flock Together? Clustering and Avoidance Patterns of Similar Retail Outlets. *Geographical Analysis* 45[2]: 123–149. CrossRef.

Krugman P (1979) Increasing returns, monopolistic competition, and international trade. *Journal of International Economics* 9[4]: 469–479. CrossRef.

Krugman P (1991) *Geography and trade.* MIT Press

Larsson JP, Öner Ö (2014) Location and co-location in retail: a probabilistic approach using geo-coded data for metropolitan retail markets. *The Annals of Regional Science* 52[2]: 385–408. CrossRef.

Lehocký F, Rusnák J (2016) Regional specialization and geographic concentration: experiences from Slovak industry. *Miscellanea Geographica – Regional Studies on Development* 20[3]: 5–13. https://www.degruyter.com/downloadpdf/j/mgrsd.2016.20.issue-3/mgrsd-2016-0011/mgrsd-2016-0011.pdf

Lessmann C (2005) Regionale Disparitäten in Deutschland und ausgesuchten OECD-Staaten im Vergleich. *ifo Dresden berichtet* 3/2005: 25–33

Lessmann C (2014) Spatial inequality and development - Is there an inverted-U relationship? *Journal of Development Economics* 106: 35–51. CrossRef.

Lessmann C (2016) Regional inequality and internal conflict. *German Economic Review* 17[2]: 157–191. CrossRef.

Lessmann C, Seidel A (2017) Regional inequality, convergence, and its determinants – a view from outer space. *European Economic Review* 92: 110–132. CrossRef.

Litzenberger T, Sternberg R (2006) Der Clusterindex – eine Methodik zur Identifizierung regionaler Cluster am Beispiel deutscher Industriebranchen. *Geographische Zeitschrift* 94[2]: 209–224

Longley PA, Goodchild MF, Maguire DJ, Rhind DW (2005) *Geographical Information Systems and Science* (2nd ed.). Wiley

Lorenz MO (1905) Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* 9[70]: 209–219. CrossRef.

Martin C (2015) Kreative Klasse 2015. Kreativität als entscheidender Faktor für wirtschaftlichen Erfolg: Entwicklungen und Ausprägungen in Deutschland. Research report. https://www.kreativ-sta.de/wp-content/uploads/2017/10/agiplan_Kreative_Klasse_2015_Studie.pdf

Midelfart-Knarvik K, Overman H, Redding S, Venables A (2000) The Location of European Industry. *European Economy - Economic Papers* 142

Moga LM, Constantin DL (2011) Specialization and Geographic Concentration of the Economic Activities in the Romanian Regions. *Journal of Applied Quantitative Methods* 6[2]: 12–21. https://pdfs.semanticscholar.org/aa9d/365d6a8ef4c3585595c8ba03fe373ab02010.pdf

Mulligan GF (2006) Logistic Population Growth in the World's Largest Cities. *Geographical Analysis* 38[4]: 344–370. CrossRef.

Mussini M (2017) Decomposing Changes in Inequality and Welfare Between EU Regions: The Roles of Population Change, Re-Ranking and Income Growth. *Social Indicators Research* 130[2]: 455–478. CrossRef.

Myrdal G (1957) *Economic theory and under-developed regions*. G. Duckworth

Nakamura R, Morrison Paul C (2009) Measuring agglomeration. In: Capello R, Nijkamp P (eds), *Handbook of Regional Growth and Development Theories*. Elgar, 305–328

Nischwitz G, Böhme R, Fortmann F (2017) Kommunale Wirtschaftsförderung in Bremen: Handlungsrahmen, Programme und Wirkungen. Schriftenreihe Institut Arbeit und Wirtschaft 23/2017. http://hdl.handle.net/10419/172756

O'Donoghue D, Gleave B (2004) A Note on Methods for Measuring Industrial Agglomeration. *Regional Studies* 38[4]: 419–427. CrossRef.

OECD (2019) OECD Territorial Reviews. Website. https://www.oecd-ilibrary.org/fr/urban-rural-and-regional-development/oecd-territorial-reviews_19900759

Palan N (2017) Konzentrations- und Ungleichheitsindizes: ein methodischer Überblick sowie ein empirischer Vergleich anhand der Textilindustrie. *Zeitschrift für Wirtschaftsgeographie* 61[3-4]: 135–155. CrossRef.

Peña Carrera L (2002) Tracing accessibility over time: two swiss case studies. Technical report. http://hdl.handle.net/2099.1/6327

Petrakos G, Psycharis Y (2016) The spatial aspects of economic crisis in Greece. *Cambridge Journal of Regions, Economy and Society* 9[1]: 137–152. CrossRef.

Planungsgruppe MWM (2009) Flächennutzungsplanung Gemeinde Wachtberg - Fachbeitrag Arbeiten. Report. http://www.wachtberg.de/imperia/md/content/cms127/gemeindeentwicklung/fnp-fb-arbeiten-24-02-2009.pdf

Pooler J (1987) Measuring geographical accessibility: a review of current approaches and problems in the use of population potentials. *Geoforum* 18[3]: 269 – 289. CrossRef.

Porter ME (1990) *The Competitive Advantage of Nations*. Free Press

Portnov BA, Felsenstein D (2005) Measures of Regional Inequality for Small Countries. In: Felsenstein D, Portnov B (eds), *Regional Disparities in Small Countries*. 47–62. CrossRef.

Portnov BA, Felsenstein D (2010) On the suitability of income inequality measures for regional analysis: Some evidence from simulation analysis and bootstrapping tests. *Socio-Economic Planning Sciences* 44[4]: 212–219. CrossRef.

Puente S (2017) Regional convergence in Spain: 1980-2015. Research report. Economic Bulletin 3/2017, Banco de Espana

R Core Team (2018a) R: A Language and Environment for Statistical Computing. Software, Vienna, Austria. https://www.R-project.org/

R Core Team (2018b) The R Manuals. Manual. https://cran.r-project.org/manuals.html

Reggiani A, Bucci P, Russo G (2011) Accessibility and Impedance Forms: Empirical Applications to the German Commuting Network. *International Regional Science Review* 34[2]: 230–252. CrossRef.

Ricardo D (1821) *On the Principles of Political Economy and Taxation* (3rd ed.). McMaster University Archive for the History of Economic Thought

Ripley BD (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability* 13[2]: 255–266. CrossRef.

RStudio Team (2016) RStudio: Integrated Development Environment for R. Software, RStudio, Inc., Boston, MA. http://www.rstudio.com/

Schmidt H (1997) *Konvergenz wachsender Volkswirtschaften. Theoretische und empirische Konzepte sowie eine Analyse der Produktivitätsniveaus westdeutscher Regionen*, Volume 152 of *Wirtschaftswissenschaftliche Beiträge*. Springer

Schönebeck C (1996) *Wirtschaftsstruktur und Regionalentwicklung : theoretische und empirische Befunde für die Bundesrepublik Deutschland*, Volume 75 of *Dortmunder Beiträge zur Raumplanung Blaue Reihe*. IRPUD

Schätzl L (2000) *Wirtschaftsgeographie 2: Empirie* (3rd ed.). Schöningh

Smith TE (2016) Notebook on Spatial Data Analysis. Technical report. http://www.seas.upenn.edu/~ese502/#notebook

Spiekermann K, Wegener M (2008) Modelle in der Raumplanung I: 4. Input-Output-Modelle. Presentation, Lecture "Modelle in der Raumplanung" WS 2008/2009. http://www.spiekermann-wegener.de/mir/pdf/MIR1_4_111108.pdf

Stark KD, Velsinger P, Bauer M, Bonny HW, Kricke J, Schwetlick D, Striedel HD (1981) *Flächenbedarfsberechnung für Gewerbe- und Industrieansiedlungsbereiche: GIFPRO*. Number 4.029 in Schriftenreihe Landes- und Stadtentwicklungsforschung des Landes Nordrhein-Westfalen. ILS, Dortmund

Statistisches Bundesamt (2008) German Classification of Economic Activities, Edition 2008. Dataset (XLS). https://www.destatis.de/DE/Methoden/Klassifikationen/GueterWirtschaftklassifikationen/klassifikationWZ08englisch.xls

Störmann W (2009) *Regionalökonomik. Theorie und Praxis*. Oldenbourg, Munich

Taylor JK, Cihon C (2004) *Statistical Techniques for Data Analysis* (2nd ed.). Taylor and Francis

Theil H (1967) *Economics and information theory*. North-Holland

Tian Z (2013) Measuring agglomeration using the standardized location quotient with a bootstrap method. *Journal of Regional Analysis and Policy* 43[2]: 186–197

Vallée D, Witte A, Brandt T, Bischof T (2012) Bedarfsberechnung für die Darstellung von Allgemeinen Siedlungsbereichen (ASB) und Gewerbe- und Industrieansiedlungsbereichen (GIB) in Regionalplänen. Research report, Staatskanzlei des Landes Nordrhein-Westfalen. https://www.wirtschaft.nrw/sites/default/files/asset/document/lep_nrw_flaechenbedarf_endbericht_endfassung_04122012.pdf

Vogiatzoglou K (2006) Increasing agglomeration or dispersion? Industrial specialization and geographic concentration in NAFTA. *Journal of Economic Integration* 21[2]: 379–396

von Neumann J, Kent RH, Bellinson HR, Hart BI (1941) The Mean Square Successive Difference. *The Annals of Mathematical Statistics* 12[2]: 153–162. CrossRef.

Weddige-Haaf K, Kool C (2017) Determinants of regional growth and convergence in Germany. Discussion paper. Discussion Paper Series 17-12, Utrecht University School of Economics

Wieland T (2019) REAT: Regional Economic Analysis Toolbox. R package version 3.0.1. Software. https://CRAN.R-project.org/package=REAT

Wieland T, Dittrich C (2016) Bestands- und Erreichbarkeitsanalyse regionaler Gesundheitseinrichtungen in der Gesundheitsregion Göttingen. Research report, Georg-August-Universität Göttingen, Geographisches Institut, Abteilung Humangeographie. http://webdoc.sub.gwdg.de/pub/mon/2016/3-wieland.pdf

Wieland T, Fuchs H (2018) Regionalökonomische Disparitäten im Spiegel von Raumtypisierungen. Ein Konzept zur Identifikation strukturell benachteiligter Gebiete in Südtirol (Italien). *Standort - Zeitschrift für Angewandte Geographie* 42[3]: 152–163. CrossRef.

Williamson JG (1965) Regional Inequality and the Process of National Development: A Description of the Patterns. *Economic Development and Cultural Change* 13[4]: 1–84

Yamamura S, Goto H (2018) Location patterns and determinants of knowledge-intensive industries in the Tokyo Metropolitan Area. *Japan Architectural Review* 1[4]: 443–456. CrossRef.

Young AT, Higgins MJ, Levy D (2008) Sigma Convergence versus Beta Convergence: Evidence from U.S. County-Level Data. *Journal of Money, Credit and Banking* 40[5]: 1083–1093. CrossRef.

Discussions

REGION
The Journal of ERSA
Powered by WU

Volume 6, Number 3, 2019, D1–D21
DOI: 10.18335/region.v6i3.281

journal homepage: region.ersa.org

The future of European communication and transportation research: a research agenda*

Cathy Macharis[1], Karst T. Geurs[2]

[1] Vrije Universiteit Brussel, Brussels, Belgium
[2] University of Twente, Enschede, The Netherlands

Received: 26 September 2019/Accepted: 13 November 2019

Abstract. Our mobility system is changing rapidly. We are at the crossroads of major changes in the way we travel and deliver goods. Research agendas are adapting to this changed environment with new challenges and opportunities. This paper presents a research agenda for the future of transportation research structured along eight cluster topics of the Network on European Communication and Transport Activities Research (NECTAR). The research agenda firstly highlights the growing complexity and need for multi- and interdisciplinary transportation research. Secondly, sustainability needs to be addressed in transportation research in its full meaning, including relationships between policy-making investigations and environmental and equity effects. Thirdly, Information and Communication Technologies (ICTs) and digitalisation, the development of autonomous vehicles, and shared mobility will have profound impacts on economies, spatial interactions all-around the world, and the availability of high resolution spatial and transportation data. Digitalisation generates many new research opportunities but also give rise to new concerns about privacy, safety, equity and public health.

1 Introduction

The field of transportation research has been growing in many different directions reflecting the wide variety of disciplinary orientations such as civil engineering, geography, urban and regional economics, mathematics, sociology, political science, psychology, computer science, health science, environmental science. This development is supported by a wide range of networks and conferences in different regions around the globe, which bring together transportation researchers. There are the well-known world conferences such as the World Conference on Transportation Research (WCTR) and the annual meetings of the Transportation Research Board (TRB) in Washington D.C., United States. Also, several disciplines have developed smaller specialized conferences, on specific sub-disciplines or specific modes. For example, rail transport, walking, cycling, or aviation.

In the late 1980s, transport problems were recognized as a European-wide phenomenon, with the growth in international trade, travel and telecommunications (Banister 1991).

*In collaboration with Moshe Givoni, Bart Jourquin, Robin Hickman, Andrew R. Goetz, Imre Keseru, Maria Attard, Edoardo Marcucci, Wafa Elias, Debbie Niemeier, Sandra Melo, Johan Woxenius, Anne Goodchild, Liv Osland, John Östh, Anette Haas, Olivier Bonin, Eric Vaz, Peter Nijkamp, Luca Zamparini, João Romão, Juan Carlos Martin, Aura Reggiani, Ahmed El-Geneidy, Benjamin Büttner, Pierre Zembri, Karen Lucas, Dick Ettema, Tanu Priya Uteng, Michael J. Widener, Luc Wismans, Emmanouil Tranos, Tuuli Toivonen and Elisabeth Mack.

Additionally, the breaking down of barriers within Europe made clear the need for a common transportation research agenda and a mechanism to include researchers from across Europe and from different disciplines. In 1992, the research network 'Network on European Communication and Transport Activities Research' (NECTAR) started as a non-profit and membership-based organization. NECTAR celebrated its 25th anniversary in 2017 in Madrid during its bi-annual conference (Geurs 2018). In the early days of NECTAR, research was significantly influenced by the process of European integration with the Maastricht Treaty in 1992, which also opened many new research questions regarding how the European Union (EU) will allow cross border transport and create new networks. This European perspective is quite clearly coming out of the initial clustering of themes, such as: 'Networks', 'Transport, Communications and Spatial Evolution in Europe', 'Transport, Spatial Opportunities and Borders' and 'Economic Analysis and Transport Policy Processes in Europe'.

The 25th NECTAR anniversary was a good moment to reflect on what changed and what should be the research priorities for the next 25 years. Also, a discussion on a transportation research agenda seems opportune and necessary given current contemporary societal problems such as global warming and rapid technological developments in the transportation field. In the 25 years after the birth of NECTAR, the transport landscape has changed dramatically. Digitalisation and new technologies are changing our travel behaviour and logistics activities.

In this discussion paper, we reflect on what the research agenda for the coming 25 years should be. Over 60 NECTAR members from a variety of disciplines have contributed to this paper. Firstly, we received inputs from 26 NECTAR members who participated in a small survey (distributed to over 200 NECTAR members) in which we asked them to identify research directions likely to emerge in the next 25 years. Secondly, the coordinators of the clusters (four co-chairs per cluster), have discussed future research directions using the survey results as inputs for their discussion. This resulted in a research agenda for the next 25 years. In the remainder of the paper we firstly describe the scope of NECTAR, followed by the history and research agenda of each of the eight clusters, and lastly provide a synthesis and conclusion.

2 The scope of NECTAR

The aim of NECTAR was (and still is) to foster research collaboration and exchange of information in the fields of transport, communication and mobility among European scholars from different disciplines and countries, with particular emphasis on a social science orientation. NECTAR members study the behaviour of individuals, groups and governments within a spatial framework. The NECTAR research community utilizes a wide variety of perspectives to analyse the challenges facing transport and communication, and the impact these challenges have on society at all levels of spatial aggregation.

So far, NECTAR has hosted 15 international conferences and organized over 100 workshops and special sessions around the world. Also, many joint cluster events have been organized on cross-cutting research areas. The clusters have collaborated in hundreds of publications. Over 30 books and special issues in international journals have been published since 1992. In the NECTAR book series, published by Edward Elgar, eight books have been published since 2011 on topics such as 'Transportation and economic development', 'Accessibility, Equity and Efficiency', 'ICT for Transport', 'Smart Transport Networks, 'City Distribution and Urban Freight Transport', and 'Transport, Space and Equity' .

The core of NECTAR consists of thematic and multidisciplinary clusters. The clusters come together frequently in between the larger bi-annual NECTAR conferences and organize workshops and special sessions on specific research themes. The clusters are dynamic in nature, and pursue new emerging research themes. In the NECTAR history, clusters have stopped (only one of the five original clusters is still active), new clusters have been formed and some clusters have shifted their focus.

Currently, NECTAR comprises of 8 clusters: 'Transport Infrastructure Impacts and Evaluation' (Cluster 1), 'Policy and Environment' (Cluster 2), 'Logistics and Freight'

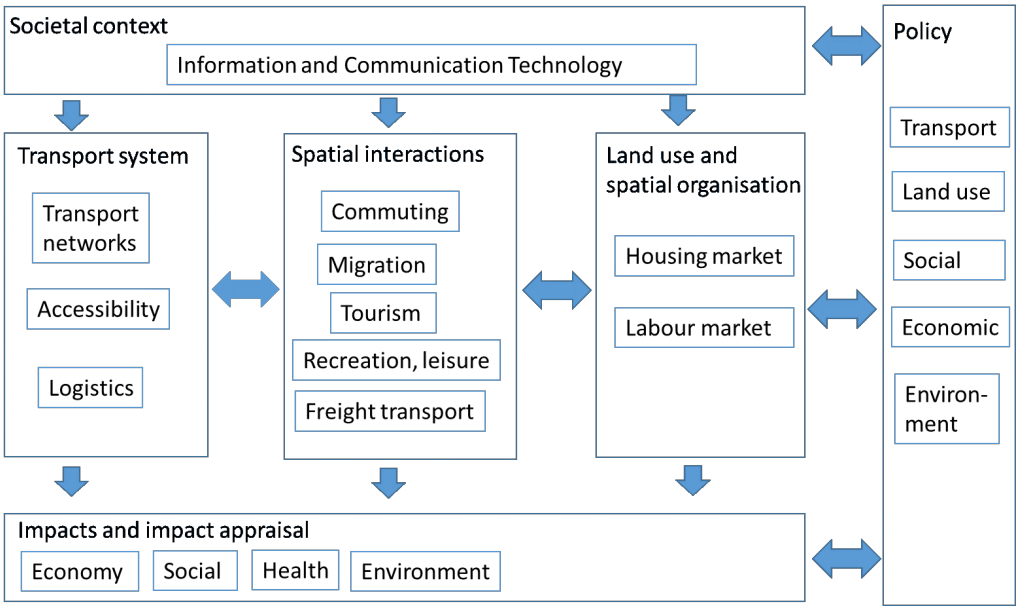


Figure 1: The scope of the NECTAR

(Cluster 3), 'Commuting, Migration, Housing and Labour Markets' (Cluster 4), 'Leisure, Recreation and Tourism' (Cluster 5), 'Accessibility' (Cluster 6), 'Social issues and Health' (Cluster 7) and 'ICT' (Cluster 8). The scope and relationships of the clusters are visualized in Figure 1. The clusters cover the reciprocal relationships between transportation (transport networks, transport and mobility providers, accessibility, logistic operators), spatial interactions (commuting, migration, tourism, freight transport) and land use (housing and labour markets). Moreover, several NECTAR clusters address governance issues and examine the impacts of transportation policies and relevant land-use, social, economic and social policies that influence the transport system, spatial interactions and land use, and their impacts on society (economy, social, health, environment).

3 Transport networks, impacts and evaluation

3.1 History and background of the cluster

The 'Networks' cluster has been active since the start of NECTAR in 1992. As noted in the introduction, research in NECTAR in the early days was very much inspired by the process of European integration allowing cross border transport and creating new networks. Throughout its history, the Networks cluster has organised over 30 workshops and special sessions. The cluster focused on the analysis and evaluation of complex transport networks, including air transport. In more recent years, emphasis is put on the social impacts of infrastructure and methodologies to assess it. In 2017, the title of the cluster changed to 'Transport Infrastructure Impacts and Evaluation'. The new title better suits the scope of research issues discussed in recent meetings and publications, such as macro level impacts such as the wider economic impacts of transport networks (Bråthen, Givoni 2017) and micro level impacts such as social inequity. Recent years saw research activities dedicated to more specific issues like: 'Exploring equity issues of ICTs for transport networks' (Thomopoulos et al. 2015); 'A future for Walking and Cycling networks'; 'Efficient European multimodal networks'; 'Smart transport network investments and smart outcomes?', and 'Enhancing network efficiency: air transport and sustainability' (Miyoshi et al. 2018). Moreover, the wide range of spatial and social equity impacts of transport systems has been explored by the cluster (Hickman et al. 2019).

3.2 A research agenda

A broader evaluation approach

Future research is needed to widen the scope of current transport policy development and project appraisal approaches. Travel time is central to our understanding of transport, and to analysis and decision making in the transport sector. Currently, transport planning is based on the rationale that all travel is 'wasteful' and travel time ought to be minimised. This is complemented by the argument that greater choice between activities at alternative destination options is beneficial. Hence, a better transport system is one that provides a wider range of destinations that can be reached in the time available to any individual. The inevitable consequence of this thinking is to promote speed as the primary objective of transport systems with a view to 'saving time' (Banister 2011). This leads to longer travel distances, has distributional outcomes (greater inequality), and results in greater use of resources, as higher speed increases energy consumption and carbon emissions. Banister et al. (2019) introduce the concept of reasonable travel time defined as the door-to-door journey time that is acceptable to the individual traveller for reaching a particular destination, and its associated activities, given the conditions provided to turn 'lost time' into 'useful time' while travelling. This means prioritising more than just travel time and speed. It can be argued that as a general rule, improving the journey experience would be an easier and cheaper means to achieve this, as compared with reducing the door-to-door travel time. Turning into the assessment of impacts, clearly this must move beyond the focus on travel time savings, to consider other various social, environmental and economic impacts and importantly the spatial impacts of (investment in) transport infrastructure, development impacts in particular, and at all spatial levels.

In this regard, the Capability Approach (Beyazit 2011) offers a new type of 'impacts' that transport infrastructure contribute to and related impacts on Mobility and Mobility Capital (Shliselberg, Givoni 2018). A different set of impacts, some of which are new, requires the development of different, new or modified, evaluation tools. The main questions here will be if measurement is possible and if so, then how to measure, and can (should) the impact be quantified and/or 'monetized'? The shifting focus in urban transport planning to walking and cycling infrastructure, which are considered 'small' compared to motorized transport infrastructure likewise require a new evaluation approach, that of so-called 'small' projects.

Digitalisation, technology and mobility services

The complexity of multimodal transport network analysis is growing as the landscape of European urban transport networks is becoming increasingly varied with the emergence of shared mobility services such as bike sharing, car sharing and ride sharing. While technology that enable automated vehicles will be transformative, it is undeniable as well that the shift from transportation systems to both Mobility as a Service (MaaS) and the sharing economy are equally transformative. Such shifts may well have deeper impacts on accessibility. The traditional dichotomy between private transportation (automobile in particular), and public transportation is quickly becoming irrelevant. Ride hailing services (UBER, LYFT, Bla-Bla-Car, etc.) as well as dockless (e-)bikes and personal light electric vehicles (PLEV), such as e-scooters, are creating a new geography that is less impeded by fixities. Research is needed to analyse and quantify the network implications and the societal impacts of these new transport modes, in particular for equity, social inclusion, and traffic safety.

Uncertainty in transport planning and network design

In a number of spheres of policy, it appears uncertainty has intensified in the face of globalisation, economic instability, climate change, technological innovation and changing consumer preferences (Lyons, Davidson 2016). There is growing interest in, and use of, techniques that can help decision making processes where deep uncertainty is involved. The 'deep uncertainty' that exist with respect to future mobility practices (see for example Givoni, Perl 2017) means also that such infrastructure must be 'flexible' to

become 'adaptive' (Haasnoot et al. 2013). Lyons, Davidson (2016) advocate a stronger orientation towards regime-testing. Through regime-testing, our understanding of the nature of the world is challenged and uncertainty is treated as an opportunity to shape the future.

4 Commuting, migration, housing and labour markets

4.1 History and background of the cluster

An important focus throughout the history of NECTAR is related to transport and spatial interactions. The NECTAR cluster 'Commuting, Migration, Housing and Labour Markets' has, since its inception, focused on topics related to commuting, migration, housing and labour markets. These topics are fundamental for the functioning of the economy and for the everyday life of many people. Haas, Osland (2014) summarise the complex relationships between commuting, migration and the housing and labour markets in an editorial of a special issue with NECTAR workshop papers in the journal of Urban Studies. In order to obtain an efficient labour market, workers should locate and relocate to where the relevant jobs are to be found. From an economic point of view, commuting and migration are essential elements for obtaining a well-functioning labour market. Finally, housing is a basic good and for many people rents or loan repayments make up a relatively large proportion of their overall budget. Due to its central role during the financial and economic crisis starting in 2007-2008, it has become clear that housing prices are important for the development of the overall economy in many countries (e.g. see Zabel 2012). Moreover, across the globe and especially in Europe, migration flows challenge regional labour and housing markets. The cluster encompasses theoretical and empirical research that focuses on analyses that increase the understanding of these linked topics, methodologically as well as for policy implications.

4.2 A research agenda

The past activities of the cluster 'Commuting, Migration, Housing and Labour Markets' has mainly focused on the complex interactions between housing and labour markets, and on patterns of migration and commuting. We believe this will also be an important research focus in the future. This focus is also echoed by the responses of the survey among NECTAR members. The overall umbrella of the research of this cluster will be the links between the various spatial interactions and different facets of sustainability, including environment, affordability, integration, participation, efficiency and fairness. Some new research directions will be pursued given a range of structural and demographic changes in the economy, while maintaining the continuum of previous research directions. These new research direction are described below.

Inequalities in the labour and housing markets

The issue of inequality has been studied for many years but regional issues of inequality will be a major topic in the future. There are growing spatial disparities in housing and labour markets in centrally versus less centrally located areas. Also, a relevant research topic is inequality in accessibility to jobs for various groups of people living in different types of regions, in particular for migrants. Research should consider differences in gender, age, socioeconomic status, culture, education and their associated occupational challenges and developments. Research questions include how specific groups can manage to change their unemployment status and what is the best way (route) out of unemployment for the various groups? Also, links between accessibility, technological advances, (affordable) housing, density, and the overall environment in urban areas need to be addressed.

Advanced spatial interaction and matching analysis

Research on commuting and migration in relation to labour and housing markets will also in the next decades need to be spatially oriented, because economic and social development happens most of all as a function of proximity. Research will need to remain

focused on spatial interactions, spatial linkages, distances, and spatial matching – both with regards to housing and labour markets. In future research, a place-based life-cycle perspective will be relevant and the focus will increasingly be on various age groups. Focus will especially be placed on the older generations, given the demographic development in many European countries. The advances in geo-computation and growing availability of high resolution spatial data (big data, and combinations of spatial approaches over time) will provide new research opportunities. Big data, machine learning algorithms and geo-referenced databases will likely become more important to study commuting, migration, labour and housing markets.

Digitalisation, commuting and labour market

Studying commuting is increasing in importance in the digital age, in particular the effects of digitalisation on firms, employees and their mobility patterns. Milakis et al. (2017, 2018) state that automated vehicles could have significant implications for cities and transport systems ranging from first order (e.g., traffic, commuting, travel demand), second-order (e.g. vehicle ownership and sharing, location choices and land use, labour and housing markets) to third-order (energy consumption, air pollution, safety, social equity, economy and public health) effects. All of these implications are highly uncertain and are important research topics. Furthermore, the shift from transportation systems to both MaaS and the sharing economy, will also impact vehicle ownership, parking demand, and the need for parking supply in cities. Furthermore, in many developed countries digitalisation has changed the labour market. The Netherlands, for example, have seen a very rapid increase in the number of solo self-employed (solopreneurs) over the last decades which has led to an increase in the demand for flexible work spaces and reduces inner-city car traffic, as most Dutch solopreneurs travel to work by bike or public transport. Self-employed workers are likely to have distinctive travel behaviour because, compared to employees, they have greater autonomy over work scheduling and are less affected by imperfect information about the labour market (Shin 2019). Studies in the United States suggest that self-employed commuters have a shorter commuting distance and time than their 'traditional' employee counterparts. However, this seems to be offset by increased travel distance and time for other work-related and non-work purposes (Shin 2019).

5 Transport, leisure, recreation and tourism

5.1 History and background of the cluster

An emerging research theme which was recently recognized within NECTAR was transportation, communication and mobility in relation to leisure, tourism and recreation. The cluster on 'Leisure, recreation and tourism' started in 2016. The research theme of this cluster is truly multi- and interdisciplinary. Research on leisure, tourism and recreation is booming in many disciplines: economics, sociology, psychology, management sciences, and environmental sciences among others. Transport constitutes a crucial yet understudied element in tourism, since it connects tourists from their origin to their destination, it creates the conditions for mobility within destinations, or to travel to secondary destinations. Moreover, in some cases, transport can be seen as a core component of the tourism experience in itself (e.g. scenic railways and buses). The interactions between transport, leisure, recreation and tourism are growing in importance as also tourism is one of the fastest growing economic sectors. Jointly, travel and tourism currently account for almost 10% of world GDP, when direct, indirect and induced effects are considered (WTTC 2014). Moreover, the increase in leisure time in many countries has constituted one of the most remarkable trends in the last decades matched by a decrease in working hours (OECD 2009). Consequently, a large amount of the travel behavior and demand is increasingly determined by ecological, social and entertainment motives (discretionary time consumption).

5.2 A research agenda

Research in this field should focus on sustainable tourism development and management, taking into consideration contemporary societal problems, such as climate change and the limits of non-renewable sources of energy. Several core and interrelated problems require research along different lines.

Sustainable tourism development

Attention to the impact of tourism on the environment, and climate change in particular, is growing since the 2000s. Research interest on climate change and tourism has increased from, on average, only 0.9% of all publications in the tourism domain in the 1990s to 2.6% in the 2000s and up to 3.4% in the 2010-2016 period (Peeters 2017). Given the current growth in tourism and the absence of interventions to reduce its climate impact, the tourism sector is likely to render the Paris (2015) climate targets unachievable. The tourism sector has pledged to reduce its greenhouse-gas (GHG) emissions, specifically 70% by 2050. However, current emission trends in the tourism sector will result in a tripling in the same timeframe (Gössling, Scott 2018). The reduction of GHGs from aviation are a crucial element to develop a credible mitigation strategy. Analysis of the environmental pressure of day-visitors and tourists to Amsterdam for example showed that long-haul tourists accounted for less than 25% of tourism revenues but were responsible for 70% of the environmental footprint of inbound tourism to Amsterdam (Peeters, Schouten 2006). For the next decades, analysis of economic efficiency (the cost competitiveness of a destination and ensuring access to a destination) and sustainable tourism development (environmental and landscape impacts) will be an important research field.

Advanced multi-scale governance analysis

Sustainable tourism development is complex as it requires institutional coordination across different territorial scales. There is a complementarity of scales between international (for mobility between origin and destination), national (for mobility between origin and destination or between the primary and secondary destinations), regional (for mobility between the primary and secondary destinations) and local (for mobility within the destination) transport services and networks. Moreover, there are many different stakeholders in different sectors (transport service providers and regulators, tourism service providers, regulators and management organizations). Romão et al. (2018) note that coordination between tourism transport services within the local territorial scale of the destination can already be problematic regarding issues such as pricing, route planning or possible conflicts between the demand of tourists for transport services to specific locations within the cities and the daily needs of the resident population. Contemporary developments in transport infrastructures (mostly in less developed countries) in the context of strong tourism development (in particular in urban areas) justify increasing efforts in research and planning.

Digitalisation and Information, Communication and Telecommunications (ICTs)

ICTs are increasingly important in order to match supply and demand of tourism services, in both space and time (e.g. Romão et al. 2015). On the other hand, increasing mediatisation and interactivity related to the development of social networks reinforce the importance of tourists as co-creators of destination images. As sustainability is a key issue for the creation of the image of contemporary tourism destinations, sustainable transport services and networks are crucially important for ecological protection and economic efficiency, but also for the image and differentiation of each destination. Moreover, ICT and digitalization also exert a deep impact on the renovation of ticketing services for transport, increasing flexibility and supporting the complementarity between different types of transports in a context of inter-modality. Thus, the deep impacts of ICT and digitalization on the provision and utilization of both tourism and transport services will offer a wide field of research opportunities in the future.

Advanced tourism data metrics and analysis

The present advances in geo-computation and spatial analysis further enhance the potential of understanding socio-economic dynamics at regional level, given the existence of large digital data sets (GPS, mobile applications, social media). In regards to these data sets, complex system analysis brings together a unique opportunity for enabling an adaptive spatial vision for tourism, leisure and recreation, as exemplified by Raun et al. (2016). Quantitative and qualitative methods may be used and intertwined with Geographic Information Systems and Science. In particular, for enhancing the understanding of the limits of carrying capacity as well as the vulnerability of fragile regions of the future. Advanced tourism data metrics and analysis become a key part of modern tourism and recreation research. Furthermore, from a statistical and modelling perspective, we observe an increasing use of (physical and virtual) network analysis, spatial autocorrelation models, econometric models, agent-based models, structural equations models, and multicriteria and benchmark studies. The increasing availability of adequate datasets as well as the development of more sophisticated user-friendly computation methods and tools will ensure that research in this broad field will tend to be even more important in the future.

6 Logistics and freight transport

6.1 History and background of the cluster

The demand for freight transport is still increasing with a fast pace. The underlying trends for this are that on the one hand of the supply chain is globalization and on the other hand there is a fragmentation of flows due to e-commerce, nanostores and just-in-time deliveries. This increase in freight transport demand poses serious questions in terms of sustainability. How can we get towards a sustainable and zero-emission deliveries of goods? In 2001, the cluster 'Logistics and Freight' started. In the early years of the cluster's research, intermodal transport was its focus (Macharis, Marcucci 2004, Caris et al. 2013). The cluster further evolved along the line of the supply chain, and in recent years focused attention on 'the last mile' and city logistics. The cluster's activities were the last ten years dedicated to more specific themes like sustainable logistics (Macharis et al. 2014), multi-perspective analysis on city distribution and urban freight transport (Macharis, Melo 2011), and the role of planning towards sustainable urban mobility (Brůhová Foltýnová et al. 2018).

6.2 A research agenda

Looking forward, two main topics are clear from the survey responses: the pressure from online shopping and the integration between freight and passenger transport. Both topics are explained below.

Digitalisation and supply chains

Digitalisation has drastically changed the traditional supply chains. The role of retailers becomes unclear and the pressure to allow for online sales options is enormous but often not implementable in a profitable way due to the increased logistics costs. For all actors it is important to implement more sustainable solutions. According to the respondents of the survey, several innovative solutions to consider are smart city logistics, drones, and autonomous vehicles. This change in supply chain characteristics will also have an impact on the location patterns of commercial establishments which needs to be researched more closely.

Integration of freight and passenger transport

Some respondents pointed to the possibilities of integrating passenger and freight transport (Arvidsson et al. 2016). This can be done in several ways. First, it can mean that facilities for passenger transport could also be used for goods: night use of parking for last mile deliveries, transport interchanges, etc. Parking and relationships between urban freight

and land use patterns should be analysed more closely. Next, it can also mean that the crowd, people like you and I will be used for crowd logistics and shipping. Recent research showed that crowd logistics is not sustainable if it is done by private cars (Buldeo Rai et al. 2017). Instead, public transport-based crowdshipping, in which only public transport or walking is allowed to bring the parcels to the end destination, represents an environmentally friendly and acceptable solution (Gatta et al. 2019, Serafini et al. 2018). Possibilities to bundle and use public transport, bike or walk should be incorporated in the concept but new types of employment raise social issues (Dablanc et al. 2017). Thirdly, it was pointed out that this integration between passenger and freight transport should be done as well at the governance level, as often freight and passenger transport is treated by different departments. Recent research showed that collaborative governance models, based on a proactive and effective stakeholders' cooperation, produce positive effects on both the environmental and service performance (Macharis et al. 2019, Marcucci et al. 2017a). Of course, attention should still be put on green logistics and more efficient freight transport. For the maritime transport, the last mile, and everything in between – it is important to have better monitoring tools of emissions and external costs. Possibilities for more sustainable solutions should be researched. The definition, implementation and up-take of effective solutions is not an easy task and there is a need for specific assessment methods.

Diffused and standardised logistic protocols

If the future of freight transport is heading towards sharing and collaboration, or even the physical internet in which information on the resources from the logistics sector would be shared in order to use the available capacity fully, protocols should be standardised. Still a lot of discussion and more fact-based evidence has to be collected for advancing knowledge on the gains and gain sharing of collective systems. Next to (ICT) technology, we will need more research on the behavioural aspect of all actors in the logistic chain, including the consumer. How is behaviour changed and what is driving this behaviour? Moreover, it is important to predict future behaviour (Marcucci, Gatta 2016) and how to stimulate behaviour change (e.g. gamification, Marcucci et al. 2018).

7 Accessibility

7.1 History and background of the cluster

It is well known that the definition of accessibility originates from Hansen (1959) as 'the potential of opportunities for interaction'. Geurs, van Wee (2004) extend this definition contending that the accessibility analysis should be based on four main components: (1) land-use; (2) the transportation system; (3) the temporal or time dimension; and (4) the individual or segmentation analysis. Geurs and van Wee extend the what (opportunities) and the how (interaction) with the when (time) and the who. There are many different approaches to measuring accessibility. The measures and methods are categorized by Martín, Reggiani (2007) into six different categories: (1) potential of opportunities; (2) physical measures; (3) expected utility; (4) inverse function of competition; (5) joint accessibility; and (6) dynamic accessibility.

The cluster 'Accessibility' started in 2008 and has become a leading group of accessibility researchers. Since the seminal workshop, 19 cluster events have been organized, and special issues and books have been organized on accessibility modelling (Martín, van Wee 2011, Reggiani, Martín 2011, Geurs et al. 2015), accessibility and planning of urban settlements (De Montis, Reggiani 2013), the impact of accessibility on social and economic activities (De Montis, Reggiani 2012), accessibility and spatial interactions (Condeço-Melhorado et al. 2015), the measurement of transport impedances in accessibility analysis (Geurs, Östh 2016), accessibility, transport planning and policy making (Geurs et al. 2012, Condeço-Melhorado, Geurs 2017) and linkages between accessibility, equity and efficiency (Geurs et al. 2016).

7.2 A research agenda

Use of big data for accessibility analysis

From the survey responses and recent cluster meetings it is clear that the trend that has been unfolding for the last three years on the use of big data for accessibility analysis is amplifying and quickly becoming part of the standard approaches for accessibility analysis and practice. ICT technologies, in the form of GPS, smartphones, credit cards, transport smart cards, social media posts, and new mobility service providers, generate a large amount of geo-located data as a valuable input for accessibility analysis. Nevertheless, Condeço-Melhorado et al. (2018) contend that despite big data already being available, its use in accessibility analysis is still in its infancy. In addition, they argue that theory on accessibility might be revisited in the light of the fast dynamics of big data and related micro-behavioural patterns.

New methodological approaches for accessibility measurement

Other type of responses focus more on new methodological approaches that extend the applicability of accessibility analysis to other fields. Of particular interest are the differences observed for some territories of the decay friction parameter (Östh et al. 2014). In this respect, the debate is still open regarding whether these differences are more based on the preferences of the citizens or on the constraints of the transport system and land use. Another interesting extension is that of the measurement of spatial spillovers (Gutiérrez et al. 2010). It is well known that transport investments affect not only the life of the residents and tax-payers but also tourists and other economic stakeholders. Therefore, it is important to take into account that transport investment can be suboptimal if these externalities are not considered in a cost-benefit analysis.

Accessibility, resilience, vulnerability and social capital

Other interesting extensions are related to the analysis of resilience, vulnerability and even social capital (García-Palomares et al. 2018, Östh et al. 2018a,b, Taylor 2017). These extensions deal with the spatial distribution of some components of accessibility measures and the relationship with the pivotal issues raised by the new fields included in the extensions. According to Caschili et al. (2015), the concepts of resilience, vulnerability, criticality and connectivity are all intertwined. Östh et al. (2018a) use other concepts to explain how spatial systems exhibit complex patterns of socio-economic development, such as (un)employment, income, mobility, ethnic composition, and urbanisation rates. They show that the resilience of spatial systems is co-determined by factors such as market proximity, land use, transport systems and socio-economic population composition.

Digitalisation, technology and mobility services

Technological developments towards autonomous vehicles and of the sharing economy/mobility as a service (MaaS) can have profound impacts on accessibility. Milakis et al. (2018) find three viewpoints in a group of accessibility experts analysing the impacts of autonomous vehicles in the four accessibility components. Viewpoint A expects that accessibility benefits stemming from autonomous vehicles will be highly uncertain, mainly because of induced travel demand that will likely cancel out travel time and cost savings in the long term. Viewpoint B anticipates that accessibility changes because of autonomous vehicles will have two opposing implications for urban form: densification of city centre and further urban sprawl. Finally, viewpoint C expects that those who can afford an autonomous vehicle will mainly enjoy its benefits, thus autonomous vehicles will have more negative than positive implications for social equity. Urban spaces may potentially become more frictionless and less centered on transportation nodes and links. In the decades to come, digitalisation and technological developments towards autonomous and shared mobility is expected to be an important research topic to study, notably quantifying accessibility and societal implications. As noted in Section 4.2, these technological

developments could have significant first, second and third order impacts on cities and transport systems and thus accessibility.

Affordability and equity issues that affect the normative pillar of the accessibility analysis

The normative pillar of the accessibility analysis regarding considerations of equity, social inclusion, inclusive urban development and transport mode preferences should leave their ideological trenches. More studies are needed to better understand the policy implications of transportation programs on specific groups, such as seniors or the unemployed. In this strand of literature, Guzman, Oviedo (2018) contend that it is necessary to use metrics that can reflect the moral dimension of inequality to assess redistributive policies that can alleviate the needs of some specific and targeted social groups which are currently worse-off.

8 Transport and ICT

8.1 8.1 History and background of the cluster

The technological changes related to ICTs have created new challenges for transportation research, the nature of which requires cross-disciplinary approaches. The NECTAR cluster 'ICT', which started in 2014, addresses these multifaceted impacts of ICT. The cluster addresses these challenges by bringing together researchers with an interest in the implications of ICT on urban and transport networks as well as in the use of big data sources for urban and transport analysis and modelling. The cluster has so far organised 4 workshops and special sessions, a common denominator was incorporating new sources of big data and using dedicated ICT applications for transport analysis (Wismans et al. 2018, Tranos, Mack 2018). The focus on this area reflected the urge of researchers to take advantage of the new opportunities that the recent abundance of data has created to better understand transportation.

8.2 A research agenda

ICTs have had and still have a transformative impact on transport systems, mobility and spatial-economic developments. ICTs directly affect space, the economy and travel behaviour by dramatically changing transaction and transportation costs. ICT developments drive the development of the emerging mobility services such as bike sharing, ride sharing, technological developments towards connected and autonomous vehicles.

Big data and transportation research

At the same time, ICTs are changing transportation research. The digital revolution has generated an abundance of digitally collected bottom-up data, the utilization of which in spatial and transport analysis has just begun. The breadth of such sources, which include anything from online social networks to passive and active crowd sourcing data, has the potential to assist researchers in better understanding spatial phenomena: from commuting to car speed analysis. More and more transport researchers are utilising such data sources, which are becoming part of the mainstream toolkit for transport analysis. A common trend is visible, demonstrating that even if traditional data is better and more reliable for the end-user, an increasing number of end-users require non-traditional data (from lower quality data sources) due to cost efficiency and data collection speed. This has resulted in the rapid development of methods and tools for transport analysis in two major streams. The first uses high level data to gain a better understanding of (aggregated) mobility patterns (e.g. using mobile phone data). The second stream makes use of dedicated applications for individual travel patterns (e.g. smart phone applications) (Wismans et al. 2018). Digital technologies are here to stay and the challenges they impose will not vanish. Thus, research will be needed to further combine and integrate data sources to allow for consistent travel behaviour knowledge on individuals as well as at aggregated levels.

Technological innovations and privacy, safety, equity and public health

The survey and cluster meetings indicate a need for research on autonomous vehicles, ethics, privacy issues, and public health. The changes that are about to take place in the transportation sector and the current innovations, including big data analytics and driverless vehicles, create new opportunities for discovering the balance between ethics/responsibility and costs/benefits. For example, current innovations can be used to help determine liability in accidents, streamline insurance pricing, motivate better driving practices, and improve safety. Yet, current innovations also give rise to new concerns about privacy, safety, equity and public health. Autonomous vehicles have implications for many public health issues beyond their potential to improve safety. These range from concerns about increasing automobile use and reduced use of healthier alternatives (e.g. biking, walking) to concerns that over-emphasizing autonomous vehicles may reduce public transport usages and divert funding from efforts to improve public transport (Schroll 2015, Milakis et al. 2017, Milakis 2019).

As noted in Section 4.2, technological developments have the capacity to directly affect transportation systems and there are possible second and third order effects on land use and spatial development, public health, etc. This increases the need for cross-disciplinary approaches, which can help tackle both the conceptual and methodological complexities. This is essential in order for the next 25 years of transport research to be even more informative than the previous 25.

9 Policy and the environment

9.1 History and background of the cluster

The cluster 'Policy and Environment' has been created in 2006 and focuses on analyzing, forecasting, measuring and discussing transport policy and its direct and indirect impacts on the environment. The cluster has organized 8 workshops on a variety of topics, including transport pricing, regulating of transport infrastructures and services, transitions towards sustainable mobility, modelling of the adoption of electric vehicles and inequalities in environmental quality. In recent years, the work of the cluster has focused on the role of instruments, individuals and institutions in the transition towards sustainable mobility (Geerlings et al. 2012) and the role of planning towards sustainable urban mobility (Brůhová Foltýnová et al. 2018).

9.2 A research agenda

Issues raised in the survey among NECTAR members and in discussions with the 'Policy and Environment' cluster were the following: firstly, the strong links between transport policy, the environment, equity and climate change; secondly, the role of stakeholder engagement in policy making; and thirdly the governance implications of technological development such as autonomous vehicles.

Transport policy, equity and climate change

The link between transport policy and environmental impact can be operationalized by correlating it directly to climate change. The inter-relationships between transport, the environment, and equity also signal the importance of understanding and improving governance. Advancing our understanding of the relationships between policymaking and environmental issues will be important to develop innovative methods and models capable of tackling the complexity of freight and passenger transport planning. Technology will play an important role in further progress toward meeting climate change goals and supporting policy. However, when it comes to transport, there has been much debate but little real action over issues of climate change, and the contribution of the transport sector to climate change is likely to increase both in relative and absolute terms. Without policy or regulation changes, transport emissions are forecasted to double over the next 30 years (IEA/OECD 2009). In addition to technological developments, there is a need to reduce levels of mobility, at least in the richer countries, and in particular air travel.

This requires new levels of coordination between all agencies and strong citizen support (Banister 2018). In addition to climate mitigation, governments and industries also need to invest in climate adaptation measures and policies to improve the resilience of transport systems.

Technology in general and digitalisation in particular have had a profound impact on economies all-around the world and this will change how people interact with one another, how they work, how they travel, and how they shop. At the same time, there is a need to reduce GHG emissions and regional/local growth in motorized transport. Concentrating on digitalized services connected to groceries is relevant and important. Research shows that online grocery stores can reduce up to 50% of their GHG emissions if service quality, monetary and environmental costs are included in vehicle routing decisions (Wygonik, Goodchild 2011). The success or failure of climate policies will depend on local government's approach to transport and land use. The examples of e-groceries and autonomous vehicles are illustrative of the complexity embedded in the social phenomena linked to, and influenced by, transport related choices and policies. This field of investigation calls for a research effort based on multi- and interdisciplinary approach focusing on the following main pillars: (1) decision-making process, (2) stakeholder engagement, (3) socio-technical analysis, (4) socio-behavioral analysis, (5) governance, (6) equity, and (7) safety.

Digitalisation, technological development and policy making

Technological innovations will have a strong effect on how vehicles are operated, owned or rented, and consequently on the environment. However, as noted in Section 3.2 there is huge uncertainty in the impacts of these innovations and there is need for tools and techniques that can help decision making processes where deep uncertainty is involved. Understanding the relationship between policy-making investigations, environmental consequences, and equity effects will be crucial. Furthermore, the role of government is changing with the digitalisation of society.

The emergence of shared mobility operators might be disruptive to public transport services that are heavily regulated by governments and transport authorities. In the past years, authorities have reacted in very different ways to Uber's tenacious and highly competitive approach to building (and destroying) markets. Cohen (2018) concludes that authorities need to form a clear understanding of the possible impacts of new mobility operators on the communities they serve. What makes Uber significant for transport authorities is that it is an aggressive competitor which, through its massive private-equity backing and business model of operating at a financial loss, it is able to undercut the conventional private-hire/taxi market in numerous locations causing Uber's rapid expansion (Cohen 2018).

Political acceptability is integral when considering the feasibility of implementing transport policies and their related issues as described above. This implies not only ensuring that the policies implemented produce net benefits to society, but also duly considering distributional effects that, among other issues, depend on the innovative and inclusive policymaking protocols developed. This complex and articulated process should rest on a deep and theoretically sound reflection. This process requires considering market structure, regulation, use of supporting/deterring policies (e.g. environmental taxes, tolls and subsidies), and complemented by new and innovative measures (e.g. cash-out policies, Evangelinos et al. 2018; off-hour deliveries, Marcucci, Gatta 2017; crowdshipping, Marcucci et al. 2017b).

Stakeholder engagement, transport modelling and decision making

Methodologically, recent trends testify to the usefulness of integrating discrete choice models and agent-based models. Such integration overcomes their respective limits and provides a thorough socio-behavioral analysis, accounting for stakeholder preferences and their interaction with respect to innovative policies (Marcucci et al. 2017c). We believe that stakeholders' acceptance of innovative transport policies will be crucial, since they are those who bear the consequences of the decisions made (Gatta, Marcucci 2014).

Even if public participation is considered important for the success of a decision-making process, in general it is not well-structured and there is a lack of methods to support group decision-making that can guide stakeholders' involvement towards thoroughly considered decisions (Le Pira et al. 2017). The gap of knowledge between methods used for technical/economic evaluation and those exploited to support stakeholder engagement is significant. Future research should deal with innovative methods and procedures that can be used for participatory decision-support systems and stakeholder analysis, both in passenger and freight transport planning. Stakeholders' reception, based on robust methods, such as stated preference analyses, represent an important building block for providing decision-makers with useful tools (Gatta, Marcucci 2016, Gatta et al. 2017, Valeri et al. 2016).

10 Social and health issues in transport

10.1 History and background of the cluster

The cluster 'Social and Health Issues' was established in Madrid, at the 2017 NECTAR conference with the remit to recognize the increasing importance of social and health inequalities arising from the transportation domain. The focus of the cluster is on emerging social research and health methodologies. The aim is to consider both the Global North and Global South as well as their inter-connections. Many disciplines engage in this broad topic area, bringing with them their own concepts, theories, methodologies and policy concerns. The aim is to draw together these disparate discourses to create opportunities for discussion, knowledge exchange, and co-production of new research in this area.

The research interests of the cluster can be categorized into three broad themes.

1. The impacts of transport systems and policies on social and health inequalities.
2. Inequalities in transport and health for socially vulnerable groups including, but not limited to, older adults, younger people, low-income population, and refugees.
3. Inequalities in access to services and activity centres as it relates to well-being and quality of life.

The cluster also aims to collaborate in the promotion of new and hybrid methodologies to identify and evaluate these broad issues and their interrelatedness.

10.2 A research agenda

The impacts of transport systems and policies on social and health inequalities

Strong links exist between social disadvantage, transport poverty, and health inequities, but these interconnections remain understudied, hidden and unacknowledged (Widener, Hatzopoulou 2016). In the same spirit, both the design and implementation of transport projects have largely been treated as isolated projects in the technical/technocratic domains. These systems have had huge social and health impacts, which are poorly understood and most often completely ignored in urban-transport planning exercises. For example, questions pertaining to exposure to air pollutants – when do they matter, which kinds are important for whom, and how does transport moderate their impacts? – need further investigation.

Inequalities in transport and health for socially vulnerable groups

For transport researchers and policy makers engaged in establishing agendas for sustainable transport planning, the 'social' dimensions of transport systems are much less well understood and articulated than economic and environmental factors. To a certain extent, the social and well-being dimension is being discussed in the Global North (Ettema et al. 2010), but this discussion is either completely absent or in its nascent stages in the Global South (see Priya Uteng, Lucas 2017). This gap is exemplified by various international development agencies that have insisted on including social impact assessments but often

fall short of providing coherent assessments. Thus, distributional impacts end up as checklist items in the toolkits approach. Though Vanclay (2002) highlighted this issue in 2002, still social appraisals are tick-box, standalone exercises conducted after the standard economic cost benefit analysis (CBA) has been completed.

Inequalities in access to services and activity centres as it relates to well-being and quality of life

While there is some overlap with the cluster on Accessibility, studies documenting access and accessibility to places and activities that promote quality of life and health are a key component to understanding the links between transportation and social and health inequalities. The core issue is that spatial systems interact with the accessibility afforded to them to create complex patterns of socioeconomic development, thereby either exacerbating or ameliorating access to opportunities (Farber et al. 2018, Östh et al. 2018a). These deviations often get expressed in terms of access to livelihood, health, education, and other opportunities that directly impact quality of life, well-being, and health. Inequalities in access should be studied for different demographic groups – young, elderly, children, women, disabled, as well as their interplay. Additionally, researchers should examine the ways different development sectors (health, education, employment, welfare, etc.) are accessed individually, and in concert with each other.

Promotion of new and hybrid methodologies

The transportation sector is at a pivotal junction in terms of both availability of 'new' kinds of data, and detailing these datasets in terms of their granularity and precision. This junction offers a chance to rectify the mistakes of focussing on aggregated results alone, at the cost of neglecting differences among demographic and social groups. There is a strong need for a robust approach to fuse disaggregated data collection and analyses based on social and health outcomes in the standard models like land use transport interaction models (LUTI), standard transport models, environmental impact assessments, etc. Further, the upcoming methods for designing integrated multimodal transport solutions have been found to be non-inclusive. New and hybrid methodologies to assess the 'inclusivity' of these solutions needs to be promoted.

11 Conclusions and synthesis

Our mobility system is changing rapidly. We are at the crossroad of major changes in the way we travel and deliver goods. Research agendas are adapting to this changed environment with new challenges and opportunities. In this paper we brought these research agendas together structured along the clusters of the NECTAR network. This resulted in bringing together the knowledge of many experts from Europe and beyond.

Many of the research topics described by the clusters are also linked to research topics which are related to other NECTAR clusters, which highlights growing complexity and growing need for multi- and interdisciplinary transportation research. In summary, there are three topics that will influence research in most of the clusters in the decades to come.

Firstly, sustainability needs to be addressed in its full meaning, including relationships between policy-making, environmental impacts, and equity effects. In particular, when it comes to transport, there has been much debate but little real action over issues of climate change, and the contribution of the transport sector to climate change is forecasted to increase both in relative and absolute terms.

Secondly, ICTs and digitalisation, the development of (shared) autonomous vehicles and shared mobility are topics that will be addressed by most of the clusters as they have a profound impact on global economies. These developments will change how people interact with one another, how they work, how they travel, and how they shop. It will also have a strong impact on how vehicles are operated, owned/rented, and consequently on the environment. Also, policy-making is about to undergo a paradigm shift linked to the driverless and shared mobility revolution. This raises major research issues related to the

governance of technological innovations as well as stakeholder and end-user involvements to achieve inclusive and equitable transport systems.

Thirdly, the advances in geo-computation and growing availability of high resolution spatial and transportation data generates many new research opportunities for theoretical, methodological and applied research. Specifically, research areas include the reciprocal relationships between transport systems, spatial interactions (including commuting, migration, tourism and freight transport) and land use as well as its impacts on society. However, at the same time there are challenges related to privacy as well as practicing open and reproducible research and development. Traditional and newly developing travel models will need be scrutinized for how they can ethically and responsibly harness the increasingly available big data sources.

Will these research lines be followed for the next 25 years? Probably not. We are only seeing some of the changes that are occurring. As researchers we will have to be open and keep track of further changes. Over the past 25 years NECTAR has been an active and productive network and will continue, in the decades to come, discussing these new research opportunities in an interdisciplinary and focused way. We warmly invite everyone to reflect and comment on this research agenda and join the discussions on the future of transportation research. Feel welcome to join any of the upcoming NECTAR events!

References

Arvidsson N, Givoni M, Woxenius J (2016) Exploring last mile synergies in passenger and freight transport. *Built Environment* 42[4]: 523–538. CrossRef.

Banister D (1991) The European science foundation's network for European communication and transportation activities research – the first five years. *Transport Reviews* 11[3]: 291–293. CrossRef.

Banister D (2011) The trilogy of distance, speed and time. *Journal of Transport Geography* 19[4]: 950–959. CrossRef.

Banister D (2018) *Inequality in Transport.* Alexandrine Press, The Farthings, Marcham, Oxfordshire, UK

Banister D, Cornet Y, Givoni M, Lyons G (2019) Reasonable travel time – the traveller's perspective. In: Hickman R, Lira BM, Givoni M, Geurs KT (eds), *A compantion to transport, space and equity.* Edward Elgar, Cheltenham (UK)/Northhampton (USA), 197–208. CrossRef.

Beyazit E (2011) Evaluating social justice in transport: Lessons to be learned from the capability approach. *Transport Reviews* 31[1]: 117–134. CrossRef.

Brůhová Foltýnová H, Attard M, Melo S (2018) Topical collection on the role of planning towards sustainable urban mobility. *European Transport Research Review* 10[2]: 38–38. CrossRef.

Bråthen S, Givoni M (2017) Editorial: The wider impacts from transport – What we know, what we still need to know and what does it mean? *Research in Transportation Economics* 63: 1–4. CrossRef.

Buldeo Rai H, Verlinde S, Merckx J, Macharis C (2017) Crowd logistics: an opportunity for more sustainable urban freight transport? *European Transport Research Review* 9[3]: 39–51. CrossRef.

Caris A, Macharis C, Janssens GK (2013) Decision support in intermodal transport: A new research agenda. *Computers in Industry* 64[2]: 105–112. CrossRef.

Caschili S, Medda FM, Reggiani A (2015) Guest editorial: Resilience of networks. *Transportation Research Part A: Policy and Practice* 81: 1–3. CrossRef.

Cohen T (2018) Being ready for the next Uber: Can local government reinvent itself? *European Transport Research Review* 10[2]: 57–57. CrossRef.

Condeço-Melhorado A, Geurs KT (2017) Topical collection on accessibility and policy making – Editorial. *European Transport Research Review* 9[33]. CrossRef.

Condeço-Melhorado A, Gutierrez J, Reggiani A (2015) *Accessibility and Spatial Interaction.* Edward Elgar Publishing, Cheltenham (UK)/Northhampton (USA). CrossRef.

Condeço-Melhorado A, Reggiani A, Gutiérrez J (2018) New data and methods in accessibility analysis. *Networks and Spatial Economics* 18: 237–240. CrossRef.

Dablanc L, Morganti E, Arvidsson N, Woxenius J, Browne M, Saidi N (2017) The rise of on-demand 'instant deliveries' in European cities. *Supply Chain Forum: An International Journal* 18[4]: 203–217. CrossRef.

De Montis A, Reggiani A (2012) Special section on accessibility and socio-economic activities: Methodological and empirical aspects. *Journal of Transport Geography* 25: 95–97. CrossRef.

De Montis A, Reggiani A (2013) Cities special section on 'analysis and planning of urban settlements: The role of accessibility'. *Cities* 30[1]: 1–3. CrossRef.

Ettema D, Gärling T, Olsson LE, Friman M (2010) Out-of-home activities, daily travel, and subjective well-being. *Transportation Research Part A: Policy and Practice* 44[9]: 723–732. CrossRef.

Evangelinos C, Tscharaktschiew S, Marcucci E, Gatta V (2018) Pricing workplace parking via cash-out: Effects on modal choice and implications for transport policy. *Transportation Research Part A: Policy and Practice* 113: 369–380. CrossRef.

Farber S, Mifsud A, Allen J, Widener MJ, Newbold KB, Moniruzzaman M (2018) Transportation barriers to Syrian newcomer participation and settlement in Durham Region. *Journal of Transport Geography* 68: 181–192. CrossRef.

García-Palomares JC, Gutiérrez J, Martín JC, Moya-Gómez B (2018) An analysis of the Spanish high capacity road network criticality. *Transportation*. CrossRef.

Gatta V, Marcucci E (2014) Urban freight transport policy changes: improving decision makers' awareness via an agent-specific approach. *Transport Policy* 36: 248–252. CrossRef.

Gatta V, Marcucci E (2016) Stakeholder-specific data acquisition and urban freight policy evaluation: evidence, implications and new suggestions. *Transport Reviews* 36[5]: 585–609. CrossRef.

Gatta V, Marcucci E, Le Pira M (2017) Smart urban freight planning process: integrating desk, living lab and modelling approaches in decision-making. *European Transport Research Review* 9[32]. CrossRef.

Gatta V, Marcucci E, Nigro M, Patella SM, Serafini S (2019) Public transport-based crowdshipping for sustainable city logistics: Assessing economic and environmental impacts. *Sustainability* 11[1]: 1–14. CrossRef.

Geerlings H, Shiftan Y, Stead D (2012) *Transition towards sustainable mobility: The role of instruments, individuals and institutions.* Routledge, Abingdon, UK; New York, USA

Geurs KT (2018) International NECTAR conference 2017. *Journal of Transport Geography* 70: 282–283. CrossRef.

Geurs KT, De Montis A, Reggiani A (2015) Recent advances and applications in accessibility modelling. *Computers, Environment and Urban Systems* 49[2015]: 82–85. CrossRef.

Geurs KT, Krizek K, Reggiani A (2012) *Accessibility Analysis and Transport Planning: Challenges for Europe and North America*. Edward Elgar, Cheltenham (UK)/Northampton (USA). CrossRef.

Geurs KT, Patuelli R, Dentinho T (2016) *Accessibility, Equity and Efficiency: Challenges for transport and public services*. Edward Elgar, Cheltenham (UK)/Northampton (USA). CrossRef.

Geurs KT, Östh J (2016) Advances in the measurement of transport impedance in accessibility modelling. *European Journal of Transport Infrastructure Research* 16[2]: 294–299

Geurs KT, van Wee B (2004) Accessibility evaluation of land-use and transport strategies: Review and research directions. *Journal of Transport Geography* 12[2]: 127–140. CrossRef.

Givoni M, Perl A (2017) Rethinking transport infrastructure planning to extend its value over time. *Journal of Planning Education and Research*. CrossRef.

Gössling S, Scott D (2018) The decarbonisation impasse: global tourism leaders' views on climate change mitigation. *Journal of Sustainable Tourism* 26[12]: 2071–2086. CrossRef.

Gutiérrez J, Condeço-Melhorado A, Martín JC (2010) Using accessibility indicators and GIS to assess spatial spillovers of transport infrastructure investment. *Journal of Transport Geography* 18[1]: 141–152. CrossRef.

Guzman LA, Oviedo D (2018) Accessibility, affordability and equity: Assessing 'pro-poor' public transport subsidies in Bogotá. *Transport Policy* 68: 37–51. CrossRef.

Haas A, Osland L (2014) Commuting, migration, housing and labour markets: Complex interactions. *Urban Studies* 51[3]: 463–476. CrossRef.

Haasnoot M, Kwakkel JH, Walker WE (2013) Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change* 23: 485–498. CrossRef.

Hansen W (1959) How accessibility shapes land use. *Journal of the American Institute of Planners* 25: 73–76. CrossRef.

Hickman R, Lira BM, Givoni M, Geurs KT (2019) *The Elgar Companion to Transport, Space and Equity*. Edward Elgar Publishing, Cheltenham (UK)/Northhampton (USA). CrossRef.

IEA/OECD – International Energy Agency and Organization of Economic Cooperation and Development (2009) Transport, energy and CO2. Paris: International energy agency

Le Pira M, Marcucci E, Gatta V, Inturri G, Ignaccolo M, Pluchino A (2017) Integrating discrete choice models and agent-based models for ex-ante evaluation of stakeholder policy acceptability in urban freight transport. *Research in Transportation Economics* 64: 13–25

Lyons G, Davidson C (2016) Guidance for transport planning and policymaking in the face of an uncertain future. *Transportation Research Part A: Policy and Practice* 88: 104–116. CrossRef.

Macharis C, Kin B, Lebeau P (2019) Multi-actor multi-criteria analysis as a tool to involve urban logistics stakeholders. In: Browne M, Behrends S, Woxenius J, Giuliano G, Holguin-Veras J (eds), *Urban Logistics – Management, policy and innovation in a rapidly changing environment*. Kogan Page, 274–292

Macharis C, Marcucci E (2004) Freight transport analysis and intermodality. *European Transport / Trasporti Europei* 25-26/VIII: 3–8

Macharis C, Melo S (2011) *City distribution and Urban freight transport: Multiple perspectives*. Edward Elgar Publishing, Cheltenham (UK)/Northhampton (USA). CrossRef.

Macharis C, Melo S, Woxenius J, van Lier T (2014) *Sustainable Logistics*. Emerald, Bingley (UK)

Marcucci E, Gatta V (2016) How good are retailers in predicting transport providers' preferences for urban freight policies? ... and vice versa? *Transportation Research Procedia* 12: 193–202. CrossRef.

Marcucci E, Gatta V (2017) Investigating the potential for off-hour deliveries in the city of Rome: Retailers' perceptions and stated reactions. *Transportation Research Part A: Policy and Practice* 102: 142–156. CrossRef.

Marcucci E, Gatta V, Le Pira M (2018) Gamification design to foster stakeholder engagement and behavior change: an application to urban freight transport. *Transportation Research Part A: Policy and Practice* 118: 119–132. CrossRef.

Marcucci E, Gatta V, Marciani M, Cossu P (2017a) Measuring the effects of an urban freight policy package defined via a collaborative governance model. *Research in Transportation Economics* 65: 3–9. CrossRef.

Marcucci E, Le Pira M, Carrocci CS, Gatta V, Pieralice E (2017b) Connected shared mobility for passengers and freight: Investigating the potential of crowdshipping in urban areas. Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE international conference. CrossRef.

Marcucci E, Le Pira M, Gatta V, Ignaccolo M, Inturri G, Pluchino A (2017c) Simulating participatory urban freight transport policy-making: Accounting for heterogeneous stakeholders' preferences and interaction effects. *Transportation Research Part E* 103: 69–86

Martín JC, Reggiani A (2007) Recent methodological developments to measure spatial interaction: synthetic accessibility indices applied to high-speed train investments. *Transport Reviews* 27[5]: 551–571. CrossRef.

Martín JC, van Wee B (2011) Guest editorial: What can we learn from accessibility modelling? *European Journal of Transport and Infrastructure Research* 11[4]: 346–349

Milakis D (2019) Long-term implications of automated vehicles: an introduction. *Transport Reviews* 39[1]: 1–8. CrossRef.

Milakis D, Kroesen M, van Wee B (2018) Implications of automated vehicles for accessibility and location choices: Evidence from an expert-based experiment. *Journal of Transport Geography* 68: 142–148. CrossRef.

Milakis D, van Arem B, van Wee B (2017) Policy and society related implications of automated driving: A review of literature and directions for future research. *Journal of Intelligent Transportation Systems* 21[4]: 324–348. CrossRef.

Miyoshi C, Mason K, Martini G (2018) Editorial: Enhancing the network efficiency: air transport and sustainability. *Journal of Air Transport Management* 69: 213–214. CrossRef.

OECD – Organization of Economic Cooperation and Development (2009) Society at a glance 2009. Oecd social indicators. paris: Oecd

Östh J, Dolciotti M, Reggiani A, Nijkamp P (2018a) Social capital, resilience and accessibility in urban systems: A study on Sweden. *Networks and Spatial Economics* 18: 313–336. CrossRef.

Östh J, Reggiani A, Galiazzo G (2014) Novel methods for the estimation of cost-distance decay in potential accessibility models. In: Condeço-Melhorado A, Reggiani A, Gutiérrez J (eds), *Accessibility and Spatial Interaction*. Edward Elgar Publishing, Cheltenham (UK)/Northhampton (USA), 15–37. CrossRef.

Östh J, Reggiani A, Nijkamp P (2018b) Resilience and accessibility of Swedish and Dutch municipalities. *Transportation* 45[4]: 1051–1073. CrossRef.

Peeters P (2017) Tourism's impact on climate change and its mitigation challenges. how can tourism become 'climatically sustainable'? PhD-dissertation, TU Delft, Delft, The Netherlands

Peeters P, Schouten F (2006) Reducing the ecological footprint of inbound tourism and transport to Amsterdam. *Journal of Sustainable Tourism* 14[2]: 157–171. CrossRef.

Priya Uteng T, Lucas K (2017) *Urban Mobilities in the Global South*. Routledge, New York. CrossRef.

Raun J, Ahas R, Tiru M (2016) Measuring tourism destinations using mobile tracking data. *Tourism Management* 57: 202–212. CrossRef.

Reggiani A, Martín JC (2011) Guest editorial: New frontiers in accessibility modelling: An introduction. *Networks and Spatial Economics* 11[4]: 577–580. CrossRef.

Romão J, Kourtit K, Neuts B, Nijkamp P (2018) The smart city as a common place for tourists and residents: A structural analysis on the determinants of urban attractiveness. *Cities* 78: 67–75. CrossRef.

Romão J, Neuts B, Nijkamp P, van Leeuwen ES (2015) Tourist loyalty and e-services: A comparison of behavioural impacts in Leipzig and Amsterdam. *Journal of Urban Technology* 22[2]: 85–101. CrossRef.

Schroll C (2015) Splitting the bill: Creating a national car insurance fund to pay for accidents in autonomous vehicles. *Northwestern University Law Review* 109[3]: 803–834

Serafini S, Nigro M, Gatta V, Marcucci E (2018) Sustainable crowdshipping using public transport: A case study evaluation in Rome. *Transportation Research Procedia* 30: 101–110. CrossRef.

Shin EJ (2019) Self-employment and travel behavior: A case study of workers in central Puget Sound. *Transport Policy* 73: 101–112. CrossRef.

Shliselberg R, Givoni M (2018) Motility as a policy objective. *Transport Reviews* 38[3]: 279–297

Taylor M (2017) *Vulnerability analysis for transportation networks*. Elsevier, Amsterdam

Thomopoulos N, Givoni M, Rietveld P (2015) *ICT for Transport: Opportunities and Threats*. Edward Elgar, Cheltenham (UK)/Northhampton (USA)

Tranos E, Mack E (2018) Big data: A new opportunity for transport geography? *Journal of Transport Geography* 76: 232–234. CrossRef.

Valeri E, Gatta V, Teobaldelli D, Polidori P, Barratt B, Fuzzi S, Kazepov Y, Sergi V, Williams M, Maione M (2016) Modelling individual preferences for environmental policy drivers: Empirical evidence of Italian lifestyle changes using a latent class approach. *Environmental Science & Policy* 65: 65–74

Vanclay F (2002) Conceptualising social impacts. *Environmental Impact Assessment Review* 22[3]: 183–211. CrossRef.

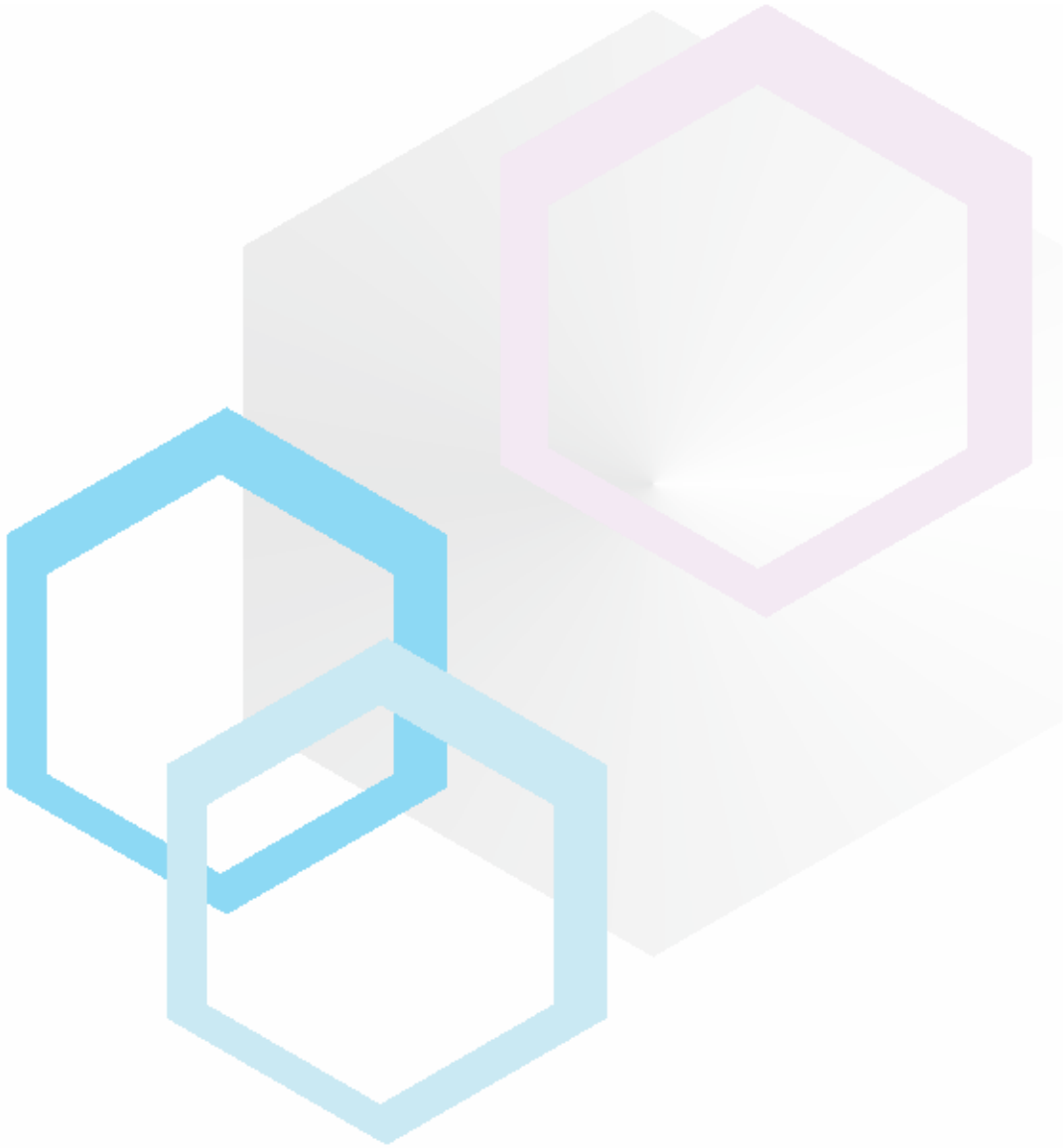
Widener MJ, Hatzopoulou M (2016) Contextualizing research on transportation and health: a systems perspective. *Journal of Transport & Health* 3[3]: 232–239

Wismans LJJ, Ahas R, Geurs KT (2018) From the guest editors: Mobile phones, travel, and transportation. *Journal of Urban Technology* 25[2]: 3–5. CrossRef.

WTTC – World Travel and Tourism Council (2014) *Travel and Tourism Economic Impact 2014*. WTTC, London

Wygonik E, Goodchild A (2011) Evaluating CO2 emissions, cost, and service quality trade-offs in an urban delivery system case study. *IATSS Research* 35[1]: 7–15. CrossRef.

Zabel JE (2012) Migration, housing market and labor market responses to employment shocks. *Journal of Urban Economics* 72: 267–284. CrossRef.



Funded by

ersa

WU

FWF