Volume 12, Number 2, 2025

REGION

The Journal of ERSA
Powered by WU

Table of Contents

Articles

Funded by
ersa WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS FWF

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

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

Articles

REGION
The Journal of ERSA
Powered by WU

Volume 12, Number 2, 2025, 1–22
DOI: 10.18335/region.v12i2.573

journal homepage: region.ersa.org
ISSN: 2409-5370

Know your competitor! Analyzing and predicting the location of competing stores: The case study of Valora at Swiss railway stations

Thomas Wieland[1]

[1] Independent researcher, Freiburg, Germany

Received: 30 October 2024/Accepted: 10 July 2025

Abstract. Location choice in retailing is a key subject of retail location theory, but is also of great practical relevance. Retail companies must assess the demand and competition situation and try to anticipate the behavior of their competitors. This study examines location choice patterns of two convenience food formats from Valora, Avec and k kiosk, at Swiss train stations. The study combines an analytical and a predictive modeling approach using econometric and AI/machine learning techniques. Possible location factors for the two formats are derived from the literature. Publicly available data from the SBB (Swiss Federal Railways) serve as the basis of the analysis. Binary logit models are built for the formats examined in order to identify the determinants of location choice. Machine learning algorithms are used to check and optimize the predictive ability of the models. It turns out that people boarding, alighting, and changing trains at train stations (which represent the main demand for convenience stores at railway stations) are an important determinant of location choice. The more frequent a train station is, the more likely it is that Avec or k kiosk will be present there. Furthermore, format-specific clustering and avoidance patterns emerge. Both Valora formats show an avoidance of each other. While Avec tends to avoid competing convenience supermarkets, this is not the case with k kiosk. With the help of machine learning, the predictive ability of the models can be greatly improved. A prediction model with high specificity and sensitivity is built for k kiosk and applied to a real case.

1 Introduction

A retailer's location choice, along with consumers' store choice, is a central topic of retail location theory, with both aspects being closely related (Reigadinha et al. 2017, Wieland 2023). Location planning is a complex process. Retail companies have been using quantitative methods in location planning for decades in order to estimate the performance of new stores (Aversa et al. 2018, Reynolds, Wood 2010). Retailers must consider demand conditions when choosing locations, and the behavior of competitors must be anticipated as well. On the one hand, a retailer competes with another retailer for the same retail space and must, for example, base its rental offer on what a competitor would pay for it. On the other hand, it must be taken into account that competitors or other non-competing stores are already located at the respective location or may expand there later. The effects of co-location with these other stores must be predicted as accurately as possible. For example, the presence of certain competitors can reduce expected sales, while others may have a positive effect. All retailers are operating under uncertainty because they do not have complete information (Orhun 2013, Zhu, Singh 2009).

Railway stations were originally purely transport hubs for people and goods. In recent decades, however, they have become important locations for retail, catering, and other services. This particularly affects large train stations that are connected to the high-speed long-distance network, regardless of whether they are centrally or peripherally located within a large city (Bills 1998, EHI Retail Institute 2023, Office of Rail and Road 2024). This is also due to the fact that mobility in European societies is increasing, particularly due to commuting, and in particular – with a short decline in the context of the Corona pandemic – train traffic is gaining (Eurostat 2024). Although transport hubs, particularly train stations, are important retail locations, they have been neglected in location research, which has mainly focused on city centers, shopping malls, and locations of supermarkets. Train stations are often seen as traffic generators for retail but are not treated as retail locations in their own right (Nilsson, Smirnov 2016, Rao, Pafka 2021). This is also problematic because several retail chains are focusing on station locations, and European national railway companies, through their own property companies, are promoting their stations as built retail destinations, similar to shopping malls (DB Station&Service AG 2017, OEBB 2024, SBB 2024b, SNCF Gares & Connexions 2024).

The *Swiss Federal Railways* (*SBB*) is the main railway company in Switzerland and has around 1,200 train stations, which are frequented by around 1.3 million travelers every day (SBB 2024c). It is also one of the largest real estate operators in Switzerland and is developing its railway stations into business centers, which makes the transport company a key player in the Swiss retail market. Commercial space in train stations is regularly advertised for retail or catering use (Neue Zürcher Zeitung 2024, SBB 2024b). The retail spaces in *SBB* train stations are highly sought after for the expansion of several food retail companies such as *Migros*, *Coop*, *Aldi*, and *Lidl*. *SBB* generally grants only fixed-term leases, which are re-tendered or renegotiated with the existing tenant after, for example, five or ten years. It often happens that the space is then taken over by a competitor (Blick 2025, Radio Frequence Jura 2024). Accordingly, there is strong competition both between retail stores in the stations and during expansion for retail space.

This study examines the location choice of the *Valora Holding AG* with regard to two convenience formats, *Avec* and *k kiosk*, at *SBB* train stations in Switzerland. *Valora* is an internationally active Swiss retail company and specializes in food convenience stores, particularly in high-frequency locations. *Valora* currently has 13 formats and around 2,800 stores in Switzerland and other European countries. In addition to *Avec* and *k kiosk*, these include catering formats such as *Caffé Spettacolo*, or bakery chains such as *Back Factory* and the Swiss branch of *Backwerk* (Valora Holding AG 2024d). *Avec* is a convenience store format with fresh to-go food and a narrow supermarket assortment that advertises with the slogan "Handmade with Love". The largest *Avec* stores in Switzerland are located in the St. Gallen (360 m2) and Andermatt (227 m2) train stations (Valora Holding AG 2019, 2024a). The kiosk format *k kiosk* advertises with the slogan "Gönn Dir was!" ("Treat yourself!") and focuses on tobacco, lottery products, snacks, and press. Originally, *k kiosk* stores were "kiosks" in the literal sense, i.e., small, often free-standing outlets in the form of a tiny house or booth that customers cannot enter. However, in recent years, more and more stores have been opened that are accessible to customers and have been expanded in terms of their (food) product range (Valora Holding AG 2024c).

This study follows a two-pronged strategy, namely an analytical and a predictive modeling approach. The first research question is of an analytical nature: *What are the determinants of the location choice of Valora convenience formats at Swiss train stations?* For this purpose, potential location factors are derived from the literature and empirically tested for their significant effect using a micro-econometric model. The second question relates to the applicability of the results for predictions: *How well can the location choice of Valora convenience formats be predicted for new situations?* For this purpose, a set of machine learning models are created, compared, and tested for their predictive ability. The study uses publicly available data sets published by *SBB*.

The paper is structured as follows. Section 2 provides an overview of the existing literature. The broad outlines of retail location theory as well as empirical studies on location choice and store performance in retailing are presented. In Section 3, independent variables of location choice with respect to the store formats examined are derived. This is followed by a description of the data sets used and the analysis and prediction models. Section 4 presents the results of the analytical model, a review of several modeling approaches in terms of their predictive ability, and an application example in which the best model is used to predict site selection. In Section 5, the conclusions are summarized and the limitations of the study are addressed.

2 Literature review

2.1 Retail location theory

Four approaches are usually attributed to retail location theory, namely (1) *central place theory*, (2) *market area models*, (3) *bid rent theory*, and (4) models of *retail agglomeration* (Reigadinha et al. 2017, Wieland 2023). The central place theory by Christaller (1933) is primarily concerned with spatial consumer behavior. Utility maximization is assumed for consumers, which means minimizing the transport costs they have to bear for a shopping trip. The demand for a specific good decreases as transport costs increase (*distance decay*). The willingness to accept traveling varies between goods depending on the frequency with which the goods are bought. The *lower range* is the minimum demand of a good that is necessary to maintain it in an economically viable manner (*demand threshold*). The *upper range* is the furthest distance up to which consumers will purchase a good offered. Profit maximization is assumed for the providers of central goods. This includes avoiding direct competitors, while clustering is assumed for suppliers of complementary goods. The result is a hierarchical system of central places of different sizes, with several goods being offered in each of these locations. Each central place has a supplementary area whose spatial extent corresponds to the upper range of the highest-ranking good offered. The theory has been formalized and extended over decades, for example, with special consideration of multi-purpose shopping (Eaton, Lipsey 1982, Ghosh 1986).

Market area models are mathematical models for calculating customer and sales flows for locations, with distance decay playing a central role. The first approaches by Reilly (1931) and Converse (1949) are deterministic and divide a market area between two locations. The probabilistic model by Huff (1962) determines the probabilities that customers from a set of origins will shop there in a system of supply locations. Consumer utility is explained by two variables that are based on microeconomic assumptions. The size of a location is regarded as an attractiveness indicator because shopping decisions are made under uncertainty, and the probability of being able to purchase the desired goods increases with the size of the location. However, *diminishing marginal utility* is assumed for the size. Consumer travel time has a non-linear negative effect on store choice because the trip to the shopping location is interpreted as *opportunity cost*. There are countless extensions to this model, e.g., to take into account the image of store chains (Stanley, Sewall 1976) or agglomeration effects (Fotheringham 1985).

The *bid rent theory* by Alonso (1964) explains urban land use and rent dynamics based on accessibility and distance from the city center. It posits that land value decreases as distance from the center increases, due to transportation costs and demand for proximity to amenities. Rent is determined by the willingness of different users to pay for location. Residential and commercial users compete for space, leading to higher rents in more desirable areas. Land near the center tends to be used for activities with higher output per area unit (e.g., retail), while outer areas are utilized for lower-value uses. The model illustrates the trade-off between land use and transportation costs, emphasizing that urban growth patterns depend on economic activities and population density.

The fourth strand of retail location theory goes back primarily to Hotelling (1929), who describes a duopoly in a linear market, where suppliers can change their location to maximize their demand. In this specific case, the best location structure for both providers is that they are located right next to each other and serve the left or right half of the market (*principle of minimum differentiation*). The influential work by Nelson (1958) stems from empirical-inductive location research. Based on customer surveys at retail locations, Nelson derives three elements of retail location success. Apart from their own attractiveness (*generative business*), the sales of retailers also depend on the attraction of compatible stores at the same location (*shared business*) and external customer frequency generators such as workplaces or public transport stops (*suscipient business*). Shared business consists, on the one hand, of the *cumulative attraction* of competitive suppliers, which arises from the fact that customers have the opportunity for *comparison shopping*, and, on the other hand, of the advantages resulting from the compatibility with other stores, which enable *multi-purpose shopping*. Nelson also derives a mathematical formula for calculating the customer exchange between two compatible retailers (*rule of retail compatibility*) and creates compatibility tables for a number of retail industries.

These topics were also dealt with early in microeconomics. Chamberlin (1933) discusses the clustering of complementary and competing stores in his seminal work towards the *theory of monopolistic competition*. Following this, it always makes sense for suppliers whose products are perfect substitutes to avoid competitors because this means they have a *monopoly on location*. In contrast, spatial clustering is favorable for suppliers of imperfect substitutes or complementary goods because the former enables comparison shopping and the latter allows for multi-purpose shopping. These considerations were later expanded and formalized with regard to incomplete consumer information (Nelson 1970, Wolinsky 1983).

Later, many of these older theories were incorporated into the microeconomic models of the "*New Economic Geography*". Here, central elements, some of which were only formulated verbally, were converted into fully mathematical equilibrium models, for example by Fujita et al. (2002) and Tabuchi, Thisse (2011).

2.2 Empirical studies regarding location choice and store network expansion

There is a heterogeneous collection of literature from economic geography and regional economics on the topic of retail location choice and store network expansion, respectively. Typically, these studies are concerned with drawing conclusions about the determinants of location choice from the empirical distribution of specific retail chains or store types. In many cases, hypotheses derived from retail location theory are empirically tested.

Larsson, Oener (2014) examine the location patterns of three retail business types in Swedish cities, especially with respect to clustering of stores. They use a geocoded database of all workplaces, and the urban areas are divided into small-scale grids. The degree of clustering is analyzed using Poisson count data models, with the number of stores in a particular industry in the grids being the dependent variable and the number of other stores and other locational variables acting as the independent variables. The presence of clothing stores is positively influenced by the presence of specialty shops and second-hand shops, while there is a negative relationship with household stores. The authors conclude that the complementarity of providers is based on the same or similar shopping frequency. It is also found that store presence is positively explained by small-scale demand (surrounding residents). Wieland (2017) applies a similar research concept to healthcare services in a rural German region. Special count data models (hurdle models) are used to investigate which location characteristics explain the number of general practitioners, psychotherapists, and pharmacies. The spatial aggregation level of the study is villages or districts. This shows that it is primarily the local demand potential that explains the number of providers examined. At the same time, clustering patterns can also be seen here; in particular, pharmacies tend to choose their location depending on the presence of medical practices.

In Canadian cities, Krider, Putler (2013) examine the location distribution of 54 retail industries and other consumer-oriented services in order to identify industry-specific clustering and avoidance patterns. Geocoded store addresses serve as the data basis to locate the individual stores. The authors use geostatistical measures to identify excessively random clustering of providers (Ripley's K, Kulldorff's D). Apart from differences between the cities, clear tendencies emerge: In particular, stores selling medium- and long-term goods (e.g., clothing, shoes, electrical goods, furniture) and specialist shops (e.g., delicatessens) tend to have a more or less strong small-area concentration. In contrast, non-specialized food retailers (e.g., supermarkets) as well as gas stations, liquor stores, pharmacies, and catering providers (e.g., ice cream shops) tend to avoid competitors.

Reigadinha et al. (2017) investigate the location structures of food retailers in a Portuguese region in order to test statements from classic retail location theory. They use point data from 273 stores that belong to eight large food retail chains. The evaluation is carried out using GIS analyses and regression models, where, among other things, store density and the distance to the nearest competitor are the dependent variables. They note a correlation between store density and population density as well as a tendency for competitors to cluster and interpret this as confirmation of the theories tested. Seong et al. (2022) examine location patterns of convenience shops in urban districts in South Korea. In their regression models, they examine the determinants of average convenience store sales and test for, among other things, the influence of convenience store density and supermarket density, while using footfall and local demand (surrounding residents and employees) as control variables. Convenience store density tends to have a positive influence on average sales, which is interpreted as a positive agglomeration effect. Supermarket density reduces sales, which can be understood as a competition effect. The footfall also increases average sales.

In a series of papers, Joseph, Kuby (2013, 2015, 2016) examine the location patterns and expansion strategies of US retail chains. Among other things, they identify that the chains have different expansion strategies and that the expansion is sometimes based on the location of the headquarters. Additionally, some chains began their expansion in large markets and continue to expand in large markets as well. The same applies vice versa to chains that initially focus on small markets. Rice et al. (2016) compare the expansion of *Walmart* and *Carrefour*. They note that *Walmart* tends to avoid competition and also serves more remote markets, while *Carrefour* primarily expands in urban areas and their metropolitan surroundings. Zhou et al. (2024) examine the expansion of Chinese electronics retailer *Suning*, counting the number of its stores at the prefecture level. They use a geographically weighted Poisson count data model. They note regionally different patterns of shrinkage and expansion. They also find that internet penetration is a predictor of the regional number of stores, in the sense that a high penetration rate can also lead to a reduction in store density.

2.3 Store performance models and predicting store sales

Another strand of literature deals with the determinants of store performance and the prediction of new store sales. The practical purpose here is to provide decision-making aids in operational location planning. Many large retail chains have been using model-based forecasts for decades (Aversa et al. 2018, Reynolds, Wood 2010). Here, regression models and/or machine learning approaches are used as well, incorporating store sales or store customer numbers as the dependent variable. Mostly, independent variables derived from retail location theory are tested empirically as well.

Possibly the first comprehensive scientific work on this topic comes from Taylor (1978), who examined the influence of location characteristics on the sales of two chains in the U.S.A., *Pizza Hut* and *Zale*. Among other things, direct competitors, population, employees, and the median income of residents in the area, as well as various aspects of micro-location, are examined as independent variables. As expected, there are positive effects from local demand and a sales-reducing effect from competition. A very early study on this is also that of Weber (1979), who examined the customer frequencies of pharmacies in a German city. Linear and intrinsically linear regression models are used to check which location factors have a significant influence. A distinction is made between locations in the city center and in the outskirts. In the first case, there is a positive influence of footfall and doctors practicing around the pharmacies, as well as a negative effect of other surrounding pharmacies. In the second case, the footfall (positive) and the distance to the nearest competitor (negative) also influence the number of customers.

Müller-Hagedorn (1991) deals with stove businesses and, unlike the previously mentioned authors, derives the location factors from the specific retail industry instead of from classic location theories. A theoretical distinction is made between, on the one hand, the situation of the consumer (e.g., level of knowledge about the products and the providers) and, on the other hand, the function of the location (time-saving or information function). Locations should therefore be evaluated differently depending on the product and type of buyer. In this case, it is argued that the information function of a location in particular makes a decisive contribution to sales, i.e., that the location should make the existence of the respective specialist store known. The result shows that both pedestrian and vehicle frequency as well as the size of the shop window correlate positively with sales.

Statistical forecasts of store performance became famous thanks to the *SLAM* (*Store Location Assessment Model*) by Simkin (1989), which – according to the author's own statement – was used for location planning after its development in several British retail chains. This is a linear regression model with, among other things, market size, accessibility, and the competitive situation of the stores as independent variables. Such models were frequently built and used in the following decades (Chang, Hsieh 2018, Themido et al. 1998). Wieland (2018) first used a panel data model to be able to take temporal effects into account; in this specific case, the yearly turnover of consumer electronics stores is investigated, with competition, regional demand, and time (as a proxy for the gaining relevance of online retailing) being the most important impacts on store performance. The topic received great attention again, especially with the emerging relevance of machine learning, which from then on was regularly used to optimize the predictive ability of such models (Ge et al. 2019, Lu et al. 2024, Ting, Jie 2022, Wang et al. 2018, Zhou et al. 2015).

Broadly speaking, almost all of these studies show positive effects of market size (e.g., residents within a travel time of X minutes) and negative effects of the presence of competitors (e.g., number of competitors in the same municipality or within a travel time of X minutes). Store characteristics are often also taken into account, e.g., store size, which typically has a positive effect on sales in terms of the store's own attractiveness. Therefore, fundamental assumptions of location theory were regularly confirmed (Turhan et al. 2013, Wieland 2018).

3 Research approach and methodology

3.1 Identification of relevant explanatory variables

In order to build a meaningful model that explains *Valora*'s choice of location in the train stations, it is necessary to derive variables from the previous location literature that can be assumed to influence the decision for or against opening. The work from location theory and empirical retail research briefly summarized in Section 2 is very heterogeneous but essentially identifies three aspects of retail locations that influence the location choice of retail companies, namely 1) the market size, i.e., local or regional demand, 2) the competitive situation, and 3) possible positive agglomeration effects due to clustering of competitive or complementary stores. However, train stations are a special type of retail location where not all of the commonly identified location factors can be directly adopted. This is mainly because train stations are transport hubs in their main function, and the retail offering there is only an "additional" service. Therefore, the essentially known location factors have to be adapted to the train station situation.

A fundamental axiom of central place theory (Christaller 1933), as well as many subsequent location theories, is that a retail store requires minimal demand to be economically viable. This minimum demand is unknown; however, one can in any case assume that the impact of demand is positive. Both studies on location selection and store performance have regularly demonstrated empirically that local demand has a positive influence on both the opening and sales of a retail store (Larsson, Oener 2014, Seong et al. 2022, Wieland 2018). Therefore, it can be expected that the probability of *Avec* and *k kiosk* being present increases the greater the demand at the station. Since the main function of a train station is that of a transport hub for loading and unloading passengers, the demand at the station is primarily determined by the number of these passengers. This also fits with the statements of Nelson (1958), for whom train stations are external frequency generators that have a positive influence on the sales of stores (*suscipient business*). Store performance studies that investigate retailers in high-frequency locations have found that footfall is a positive driver of sales in different retail industries, including convenience stores (Müller-Hagedorn 1991, Seong et al. 2022, Weber 1979). *Valora* itself explicitly states that it is looking for new retail spaces "in a highly frequented location" (Valora Holding AG 2024b). Train passengers may also be considered as a proxy variable for the (unknown) footfall within the train stations. Therefore, the average daily number of passengers is used as the independent variable of demand volume.

The aspects of the competition and positive agglomeration effects cannot be clearly separated from one another, since a competitive store may certainly increase the competitive pressure or, on the contrary, can even increase frequency. Retail location theory describes both clustering and avoidance strategies of competing retailers when choosing a location as well as positive agglomeration effects due to clustering with competitive and/or complementary stores (see Section 2.1). Nelson (1958) in particular regards competitive stores partly as a source of frequency (positive agglomeration effects) and partly as damaging to business because they increase competitive pressure. Which of the two effects predominates depends on the retail industry under consideration and cannot be determined a priori. Empirical location studies also find effects of clustering with competitors or other providers, although the effect is very industry-specific (see Sections 2.2 and 2.3). For example, Seong et al. (2022) find with regard to convenience stores (which most closely corresponds to the case examined here) that their small-scale density has a positive effect on average sales, while their proximity to supermarkets has a negative effect. However, Krider, Putler (2013) find that food retailers tend to avoid competition. In train stations, the focus of the retail offering is usually on to-go food and groceries. Other retail chains that are often present at train stations include *Coop* and *Migros* with different convenience formats. There are also other kiosks, bakeries (including those of *Valora*), and fast food and takeaway restaurants. All of these types of offerings mentioned can in principle be considered as competitors for the two *Valora* formats examined, especially for *Avec*, which is a convenience supermarket, similar to *Pronto* (*Coop*) or *Migrolino* (*Migros*), for example. However, that doesn't mean that their presence will necessarily stop *Valora* from opening a store there. Instead, *Valora* may have a specific clustering and avoidance strategy, which is, however, unknown to the public. Thus, it is in no way clear a priori which of these competitors will have a positive or negative effect on *Valora* setting up at a train station. Therefore, the numbers of all mentioned competitors are taken into account as independent variables. It is expected that at least the presence of direct competitors – especially other convenience stores like *Pronto* or *Migrolino* – will reduce the likelihood of locating in a given micro-location.

Furthermore, *Avec* and *k kiosk* themselves also compete with each other to a certain extent. Since *k kiosk* stores tend to expand further and offer more food products (Valora Holding AG 2024c), it is to be expected that *Valora* will coordinate the location planning of these two formats in order to avoid self-cannibalization. It is therefore expected that the presence of *k kiosk* at a micro-location decreases the probability of opening *Avec*, and vice versa.

In addition, other characteristics of the micro-locations in the *SBB* train stations must be taken into account. For example, the passenger frequencies are only available at the level of the entire station (see Section 3.2). However, especially in large train stations with many micro-locations (e.g., waiting hall, platform area, secondary entrances), it cannot be assumed that the frequency is the same everywhere. Therefore, other available attributes of the respective micro-location (floor, type of micro-location) are included as explanatory variables. Furthermore, the number of ticket machines at the micro-location is taken into account as an independent variable, as it can be assumed that this represents an indication of the frequency at the micro-location.

3.2 Data collection and preprocessing

Two freely available data sets published by the *SBB* were used. The first data set contains all commercial space uses in Swiss train stations, including the name, the associated train station (marked with name and unique identification code `BPUIC`) and other information (SBB 2024d). For the stores, the dataset contains, among other things, their name or chain (`Name`), a categorization (`category`, e.g., `shopping`, `sbb_services`) and subcategorization (`subcategory`; the `shopping` category includes, for example, `food`, `bakery`, or `kiosk`), the micro-location within the train station (`location_details_en`, e.g., city level, hall XY, underpass XY, floor XY), the level of the train station in which the store is located (`Ebene`), and the opening hours (`openinghours`). Ticket machines also have their own entry in the classification. The data set has 5,254 entries (download from April 3, 2024).

The second data set contains the passenger frequencies of the SBB stations (SBB 2024a). There is data on those boarding and alighting at the train stations (train passengers only), namely the annual average daily traffic (`DTV_TJM_TGM`), the annual average weekday traffic (`DWV_TMJO_TFM`) and the average non-weekday traffic (`DNWV_TMJNO_TMGNL`). This data set contains the information mentioned for the years 2018 (1,160 stations), 2022 (1,161 stations), and 2023 (1,159 stations). The data also includes the station's unique identification code (`UIC`) and information about which data from which transport companies was included in the numbers. This data set is updated annually (download from June 14, 2024). In some cases, the counts do not include passenger numbers from specific railway companies, especially those from *RBS* (*Regionalverkehr Bern-Solothurn*).

The 1,159 train stations were divided based on their micro-locations recorded in the variable `location_details_en` in the *SBB* data set, which is the basis of the analysis (n=1,443). Most railway stations only consist of one micro-location, although the large stations (e.g., Zurich, Bern, and Basel) are divided into up to 30 micro-locations on several floors. The stores were summed up by their name and their subcategory at the level of train station micro-locations. At the time of data collection, 264 *Avec* or *k kiosk* shops were located at *SBB* train stations, of which 127 were *Avec* and 137 were *k kiosk*. At least one of both formats can be found at 249 micro-locations in 203 *SBB* railway stations. The highest density of *Valora* convenience stores is with 14 stores (1 x *Avec*, 13 x *k kiosk*) at Zürich main station (average daily passenger frequency 2023: 398,300) at ten different micro-locations.

3.3 Analytical model

In the first step, a binary logit model is used for the microeconometric analysis of the determinants of location choice. The dependent variable in each model is coded in binary for the respective chain examined. It is equal to one if at least one store of the respective chain (*Avec*, *k kiosk*, and both) is present at a micro-location, and zero otherwise. The following representation of the model is based on that in Cameron, Trivedi (2005) and Greene (2012). The target variable of a binary logit model is the probability that the examined condition is true or not, which is derived from the empirical distribution of positive (1) and negative (0) events, taking into account the explanatory variables. Here the target variable is the probability that the respective chain is present at the respective micro-location, which means that the number of stores of the chain c at the micro-location m in the train station s is greater than zero:

$$\Pr(Y_{cms} > 0|\mathbf{X}_s) = p_{cms} = \frac{\exp(\boldsymbol{\beta}\mathbf{X}_{ms})}{1+\exp(\boldsymbol{\beta}\mathbf{X}_{ms})}$$

where Y_{cms} is the number of stores belonging to chain c at micro-location m in railway station s, p_{cms} is the probability that chain c is located at micro-location m in railway station s, $\mathbf{X}_{ms}$ is a set of explanatory variables (attributes of micro-location m and/or railway station s), and $\boldsymbol{\beta}$ is a set of corresponding regression coefficients.

Exponentiating both sides leads to the odds (ratio of the probability that the event occurs to the probability that the respective event does not occur): $p_{cms}/(1-p_{cms}) = \exp(\boldsymbol{\beta}\mathbf{X}_{ms})$. The logit (log-odds) equals the linear combination of parameters and independent variables: $\ln p_{cms}(1-p_{cms}) = \boldsymbol{\beta}\mathbf{X}_{ms}$. The coefficient of independent variable x_n, β_n, may also be interpreted as (semi-)elasticity with respect to the odds. The marginal effect (probability change due to a one-unit change in independent variable x_n) is the partial derivative with respect to x_n: $\partial p_{cms}/\partial x_n = p_{cms}(1-p_{cms})\beta_n$.

Table 1 shows the independent variables of the model analysis. The passenger frequency was transformed with the natural logarithm in order to achieve an approximately normal distribution and to be able to interpret the associated model coefficients as elasticity. The levels of the station and the types of micro-locations have been converted into a simplified classification with three categories each. Competitors from the same company

Table 1: Independent variables in the model analysis

Variable name	Description
	Micro-location characteristics
DTV_TJM_TGM	Station frequency (average daily boarding and alighting 2023)
level_cat	Floor in the train station (categorized) Cat 1 First floor or above 2 First basement floor or below 0 Ground floor (city level)*
microlocation_cat	Type of micro-location (categorized) Cat 1 Underpass or pedestrian bridge 2 Shopping street, gallery, passage, shopping center 0 All others*
	Competitors and other suppliers
K_Kiosk_count	No. of *k kiosk* stores at the micro-location (*Avec* model)
Avec_count	No. of *Avec* and *Avec express* stores at the micro-location (*k kiosk* model)
Migros_all_count	No. of *Migros* and *Migrolino* stores at the micro-location (Eating and drinking formats from *Migros* such as *Migros Daily*, *Migros Eatery*, *Migros Restaurant*, and *Migros Take Away* are not included; these are included in the variable *catering_count*)
Coop_all_count	No. of *Coop*, *Coop to go*, and *Coop Pronto* stores at the micro-location (Other *Coop* food formats such as *Karma* and *Sapori d'Italia* are not included, but belong to *Food_other_count* or *catering_count*, depending on the classification)
Discounter_count	No. of *Lidl*, *Aldi*, and *Spar* stores at the micro-location
Kiosk_other_count	No. of other stores of type "kiosk" at the micro-location except *k kiosk*
Food_other_count	No. of other stores of subcategory "bakery", "beverages", "butcher", "food", and "supermarket" at the micro-location except those mentioned above
catering_count	No. of other stores of subcategory "bar", "cafe", "fast food", "restaurant", and "take away" at the micro-location
vend_machine_count	No. of vending machines at the micro-location

Note: * Reference category

(e.g., different *Migros* convenience formats) were grouped together so that the distribution of these independent variables is less skewed. The linear combination of the independent variables and their empirically determined coefficients is thus:

$$\begin{aligned}\mathbf{X}_{ms} \quad = \quad & \alpha + \beta \ln \mathrm{DTV\_TJM\_TGM} + \gamma_m \sum_{m=1}^{M} \mathrm{level\_cat}_{ms} + \\ & \delta_n \sum_{n=1}^{N} \mathrm{microlocation\_cat}_{ms} + \lambda_c \sum_{c=1}^{C} \mathrm{Comp}_{ms}\end{aligned}$$

Binary logit models are estimated using the maximum likelihood method. The log-likelihood in this case is:

$$LL = \sum_{i=1}^{n} y_i \ln f(\boldsymbol{\beta}\mathbf{X}_{ms}) + (1 - y_i) \ln(1 - f(\boldsymbol{\beta}\mathbf{X}_{ms})) \tag{1}$$

where y_i is the i-th observation and n is the number of observations.

The iteratively reweighted least squares (IWLS) algorithm is used for the estimation. The significance level was set to 90%. The analysis was conducted in *R* version 4.4.0 (R Core Team 2024), including the help of the *stargazer* package (Hlavac 2022).

3.4 *Optimization of predictive ability*

The second step, after the microeconometric analysis, is about optimizing the predictive ability of the model using machine learning techniques. Machine learning (ML) is a subset of artificial intelligence (AI) and enables systems to learn from data and improve over time without being explicitly programmed for each specific task. From an ML perspective, this is a (binary) classification problem (Boehmke, Greenwell 2020, Kuhn 2008). Such questions arise in very different disciplines, for example, in banking with regard to the probability of loan default or in the medical context when predicting possible complications after treatments or assessing the risk of death. In these cases, different ML modeling approaches are used and compared with each other in terms of the accuracy of their predictions (Celio Di Cellio Dias et al. 2018, Omar et al. 2024, Shahidi et al. 2023). Here, five ML algorithms are implemented. Four of them are ensemble methods, which means that they combine a given number of (weak) learners into one aggregated learner with high accuracy, sometimes referred to as the "wisdom of the crowd" effect (Boehmke, Greenwell 2020):

1. *Decision tree* (DT): The tree consists of nodes (decisions, more precisely divisions of the independent variables) and leaves (predictions). Each node divides the data based on the input features, creating a hierarchical structure. During the training process, the tree is built by dividing the data into different groups (for categorical independent variables) or intervals (for continuous independent variables) based on the input features. Unlike the binary logit model, modeling the relationship between a dependent variable and the independent variables using one (or more) decision tree(s) is a non-parametric algorithm. The division of data in the classification tree is done using Gini impurity, which is an indicator of how mixed a node is in terms of categories:

   $$Gini = 1 - \sum_{c=1}^{C} (p_c)^2$$

   where C is the number of classes and p_c is the proportion of class c.

   The algorithm looks for the split that leads to nodes that are mixed as little as possible; that is, the Gini impurity is minimized. A decision tree may, but not necessarily, contain all explanatory variables.

2. *Decision trees with bagging* (DTBG): The second model is an ensemble method that combines decision trees and bagging (bootstrap aggregating). In this algorithm, t decision trees are formed, always with a different bootstrap sample from the data. These individual models are combined into one prediction by averaging the estimated class probabilities together. The minimum number of observations that must be present in a node for this node to be further split was set to $Split_{min} = 2$.

3. *Random forest* (RF): A random forest algorithm is an extension of bagged decision trees. A further random component is implemented here, namely only a randomly selected subset of the explanatory variables, m_{try}, is implemented for each split (here $m_{try} = \sqrt{p}$, with p being the number of explanatory variables). Bagging and random forest algorithms were tested with $t = 20$, 50, and 100 trees.

4. *Gradient boosted logit* and *gradient boosted trees*: Gradient boosting (GB) can be used for various basic models and is also an ensemble method. Here, new learners are added sequentially to the ensemble; namely, in each step i a new learner is added who is specifically fit to address the errors (residuals) of the previous one as well as possible. More precisely, this means that the algorithm iteratively adjusts the predictions to minimize a specific loss function. Here, the logistic loss function (Log loss) is used, which is a normalization of the log likelihood (Equation 1) and the default metric for binary classification problems in the used estimation package:

   $$Logloss = -\frac{1}{n} LL \tag{2}$$

   where n is the number of observations.

   The result is an aggregate of the learners created over I iterations. The fit to the overall data set usually improves with each iteration, but this does not necessarily apply to out-of-sample fitting (overfitting). The algorithm is used with both binary logit models (BLGB) and decision trees (DTGB). Both gradient boosting algorithms were tested with $I = 100$ and 200 iterations.

5. *Artificial neural network* (ANN): Artificial neural networks imitate the structure of biological neurons in brains. They consist of input, output, and at least one hidden layer, where each node (neuron) computes a weighted sum of its inputs (original input features), applies an activation function, and passes the result forward. During training, backpropagation and gradient descent minimize the loss function by updating the weights. In the present binary case, the loss function is also log loss (see Equation 2). The ANN algorithm is applied using default values (5 neurons, decay parameter equal to 0.1 for regularization) and is tested with 100 and 200 iterations.

In line with other ML classification problems, the models are assessed by a confusion matrix, which includes two performance indicators that have been used for decades to check the quality of diagnostic tests in medicine, *specificity* and *sensitivity*. Sensitivity is the share of true positives that are correctly predicted by the models. Specificity is the share of true negatives that are correctly predicted by the models (Altman, Bland 1994, Boehmke, Greenwell 2020, Trevethan 2017):

$$Sensitivity = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$Specificity = \frac{TN}{TN + FP} = \frac{TN}{N}$$

where TP is the number of true positives (number of micro-locations where $Y_{cms} > 0$ and $p_{cms} > 0.5$), TN is the number of true negatives (number of micro-locations where $Y_{cms} = 0$ and $p_{cms} < 0.5$), FN is the number of false negatives (number of micro-locations where $Y_{cms} > 0$ but $p_{cms} < 0.5$), FP is the number of false positives (number of micro-locations where $Y_{cms} = 0$ but $p_{cms} > 0.5$), and P and N are the numbers of total positives and negatives, respectively.

Sensitivity therefore makes a statement about how well the individual model predicts the cases in which a store is actually located in a micro-location. Specificity, on the other hand, documents how well the individual model predicts the cases in which there is no store. The metrics are calculated for the models based on the training data for the test data set (out-of-sample). Additionally, the $ROC - AUC$ (*Receiver Operating Characteristic - Area Under the Curve*) metric is calculated, which represents a trade-off between sensitivity and specificity. The ROC curve plots sensitivity against $1 - specificity$ at various thresholds, and the AUC measures the overall ability of the model to distinguish between classes. A higher AUC value indicates better model performance, with values closer to 1 signifying high sensitivity and specificity, while values closer to 0.5 suggest a random classification. Since in the dataset the case of an outcome of 1 is less likely than an outcome of 0 (the dependent variable is skewed), sensitivity is used as a metric for model selection. The division into training and test data sets is 90 to 10%. Model validation is undergone with 10-fold repeated cross validation with five repeats (Boehmke, Greenwell 2020). The analysis was conducted in *R* version 4.4.0 (R Core Team 2024) using the package *caret* (Kuhn 2008) and related packages such as *randomForest* (Liaw, Wiener 2002) and *nnet* (Venables, Ripley 2002), as well as own functions.

4 Results

4.1 Analytical model: Determinants of location choice

Table 2 shows the results of three binary logit models, with the presence of *Avec*, *k kiosk* or at least one of the two acting as the dependent variable (1=yes, 0=no).

The daily passenger frequency has a significant and positive influence on the probability of choosing a location at the respective micro-location, which is similar in all three models. A 1% increase in passenger frequency increases the odds of opening a *Valora* convenience store by approximately 0.4% (*Avec*: 0.364, *k kiosk*: 0.406, both: 0.434). This is not

Table 2: Binary logit model results

	Dependent variable:		
	Avec	K_Kiosk	Valora
	(1)	(2)	(3)
log_DTV_TJM_TGM	0.364***	0.406***	0.434***
	(0.051)	(0.101)	(0.054)
level_cat1	−2.667***	−0.215	−1.648***
	(0.926)	(0.689)	(0.520)
level_cat2	−16.238	−1.436	−2.289***
	(527.091)	(0.944)	(0.686)
microlocation_cat1	0.855	0.036	−0.304
	(1.114)	(0.773)	(0.668)
microlocation_cat2	0.444	2.162**	1.251
	(1.397)	(1.094)	(0.895)
K_Kiosk_count	−1.576***		
	(0.527)		
Avec_count		−1.469**	
		(0.612)	
Migros_all_count	−1.683	1.214**	0.099
	(1.047)	(0.576)	(0.453)
Coop_all_count	−0.924	3.991***	2.264***
	(0.652)	(0.627)	(0.547)
Discounter_count	−16.097	1.061	0.774
	(1,909.183)	(5.341)	(3.634)
Kiosk_other_count	−15.403	1.194	−1.053
	(919.552)	(1.153)	(1.073)
Food_other_count	0.054	−4.434***	−2.260***
	(0.320)	(0.400)	(0.273)
catering_count	0.326**	2.321***	1.373***
	(0.147)	(0.250)	(0.171)
vend_machine_count	0.842***	0.954***	1.308***
	(0.166)	(0.243)	(0.202)
Constant	−5.583***	−8.443***	−6.462***
	(0.452)	(0.958)	(0.512)
Observations	1,443	1,443	1,443
Log Likelihood	−368.285	−130.503	−404.544
Akaike Inf. Crit.	764.571	289.007	835.088

Note: *p<0.1; **p<0.05; ***p<0.01

surprising, as any store requires a minimum level of demand, and in this case this is represented by the passenger frequencies. Empirical location research with regard to stores at frequent locations has shown that footfall is a determinant of store sales (Müller-Hagedorn 1991, Weber 1979, Seong et al. 2022). However, it must be taken into account that studies that include pedestrian frequencies in their models may have an endogeneity problem. It cannot be clearly clarified which part of the frequency explains the sales or the number of customers of the stores examined and which part of it is caused *by* these stores. The direction of the causal relationship cannot therefore be fully explained (*chicken and egg problem*). However, this problem does not exist in the current case of train stations because we consider the train passengers and not the footfall (which is not available). It is plausible to assume that those boarding, leaving, or changing trains are moving from an origin to a destination (e.g., work) while using a train and typically do not take the train to a train station just to buy groceries there. That's why train stations act as an external frequency generator, which is, so to speak, a "model example" of *suscipient business* in the sense of Nelson (1958).

A minimum demand in the sense of a minimum necessary passenger frequency can also be implicitly derived from the empirical data. The train station with the lowest frequency, where *Avec* is located, has a daily passenger volume of 540 people (Murgenthal), closely followed by Flums (560 people) and Aarberg (580 people). In all three cases mentioned there is no competition (*Coop*, *Migros*, or other food or kiosk) located at the station. The smallest train station with a *k kiosk* store has 640 passengers daily and also no competition (Saanen).

The type of micro-location within the train station also has an influence on the probability of location choice, although not every characteristic is statistically significant. Both *Avec* and *k kiosk* tend to be located less often on the upper or lower floors (level category 1 or 2) of train stations. The *k kiosk* model also shows that this format is found significantly more frequently in micro-location category 2 (shopping streets, etc.). If the micro-location is a shopping street or something similar, this increases the odds of a *k kiosk* being present by a factor of exp(2.162), which equals approximately 8.688. These results can be explained by the fact that footfall is, of course, not evenly distributed within a railway station. The passenger frequencies are only collected for the entire station. It is very likely that there will be a higher frequency on the ground floor of a station than on the upper or lower floors. The same applies to shopping streets, etc., within the train stations, where many other shops are located, which in turn brings frequency.

For every *k kiosk* store that is present at the micro-location, the chance that an *Avec* store will also be located there reduces by the factor exp(-1.576), which equals approximately 0.207, all other things being equal. The reverse effect is very similar. Each *Avec* store reduces the chance that a *k kiosk* will be opened at a micro-location by a factor of exp(-1.469) $\approx$ 0.230. Since these two formats do not occupy the same sales area sizes, it is very likely that an avoidance strategy is being deliberately pursued in order to prevent internal competition (self-cannibalization). This is also plausible because *k kiosk* stores have expanded their assortment in recent years (Valora Holding AG 2024c), and it is to be expected that the markets that serve these two formats will overlap to a considerable extent. However, there are no explicit statements from the *Valora Group* that the two formats are deliberately localized according to the principle of avoidance. That this is the case remains a reasonable assumption but cannot be directly proven.

When it comes to clustering and avoidance strategies with regard to other competitors at train stations, there are apparently differences between the two formats examined. The coefficients of the variables for the presence of *Coop* and *Migros* stores as well as discounters and kiosks are all negative in the case of *Avec*, but not statistically significant. This result may also be due to the fact that *Avec* tends to compete with the above-mentioned competitors for sales space of similar size and that the *SBB* only awards the contract to one of the two if only one store is available. The format *k kiosk*, on the other hand, occupies much smaller selling spaces (see Section 1). In the latter case, the presence of *Coop* or *Migros* seems to increase the probability that *k kiosk* is located at a micro-location. Avoidance patterns of food competitors, which have been shown by, e.g., Krider, Putler (2013), cannot be confirmed in this specific case. However, the result of Seong et al. (2022), who found positive sales effects of clustering food convenience stores, may be confirmed here. Convenience supermarkets from competing companies may also act as external frequency generators. Gastronomic providers and *SBB* vending machines increase this probability in both cases. In the latter case, this is probably because the presence of ticket machines is a proxy variable for the frequency in the respective part of the railway station (see Section 3.1).

4.2 Evaluation of external model validity

Table 3 presents the three metrics sensitivity, specificity, and ROC-AUC for the three models (out-of-sample with a test dataset of 10% of all cases). The analytical model (binary logit) acts as a baseline against which the machine learning models are assessed. First of all, it can be seen in general that, as expected, the ML algorithms perform significantly better than the binary logit model. These results are consistent with those of other ML applications for (binary) classification problems (Celio Di Cellio Dias et al. 2018, Omar et al. 2024, Shahidi et al. 2023). This is not surprising since a binary logit model is not estimated with the aim of optimal predictive ability, while ML algorithms are designed for exactly that. This applies in particular to sensitivity, i.e., the correct prediction of the true positives, in this case the micro-locations where a *Valora* convenience format is actually located. In the *Avec* case in particular, a significant improvement in this metric can be seen, as here the binary logit model correctly predicts only 1.6% of the positive cases, which makes this analytical model completely unsuitable for a forecast.

The specificity is close to 100% for all models, meaning that all models almost perfectly predict the micro-locations where no *Valora* convenience format is present. However, this can be explained by the distribution of positive and negative events in the data. An *Avec* is located in only 8.8% of the micro-locations and a *k kiosk* in 9.5%, while one of the two occurs in 17.3% of all cases. Thus, at most of the 1,443 micro-locations there is neither an *Avec* nor a *k kiosk* store, i.e., the expression “no” or 0 is the most common case. A model that always predicts “no” would therefore in 91.2% or 90.5% or 82.7% of the cases make a correct prediction. This is the so-called *no information rate* that should be taken into account when assessing the accuracy of binary outcome models (Kuhn 2008). In this case, specificity cannot be used meaningfully for a comparison of predictive ability. Thus, the best model for each case is chosen with respect to sensitivity (marked bold in Table 3).

In the *Avec* case, a bagged model with 20 decision trees leads to the highest sensitivity value, with 31.6% of the true positives being predicted correctly. However, even the hit rate of less than a third can still be described as rather weak, so for practical purposes it probably wouldn’t make sense to use this model for forecasting. The random forest models with 50 or 100 trees achieve almost the same sensitivity (30.5 and 31.4%). In contrast, gradient boosted trees have slightly higher specificity but lower sensitivity. Both artificial neural networks produce lower sensitivity values compared to the tree-based models. However, the trade-off between sensitivity and specificity in terms of ROC-AUC is the highest for the ANN results. When predicting the *k kiosk* locations, the highest sensitivity of 84.7% is achieved with the decision tree bagging models with 50 or 100 trees. In this case, the model with the smaller number of trees given the same performance is considered the best model. The sensitivity of gradient boosting and ANN models is slightly lower. In the case of predicting one of the two *Valora* chains, the best performance in terms of sensitivity is achieved by the gradient boosted logit model approach with 100 iterations (67.7%). However, the specificity is the lowest of all model variants. In the last two cases, the ANN models also provide a (sometimes much) better balance between sensitivity and specificity, but not the highest sensitivity, which serves as a selection criterion here.

It turns out that, depending on the respective case, different ML model approaches lead to the best result and that a higher number of learners does not necessarily lead to a higher predictive ability. This is usually because the trained model reflects a lot of the variance in the training data, which at the same time reduces the external validity (*overfitting*) (Boehmke, Greenwell 2020). When searching for the best model, it is therefore necessary to test different algorithms with different configurations, using a performance metric that is suitable for the case at hand (in this application this is sensitivity; see above). Regarding sensitivity and specificity, it should also be said that, normally, a trade-off between the two must be made because, in any test (or model), an optimization of one indicator leads to a deterioration of the other indicator (Trevethan 2017). In the present case, this effect hardly occurs because the specificity is automatically very high, as a negative result is much more common in the empirical data. If sensitivity would not have been the decisive factor in this case, an ANN approach would have been the best model in all three cases.

4.3 Model simulation: Which station commercial spaces are suitable for k kiosks*?*

The model analysis is now used for a practical case: The *SBB* permanently advertises commercial space in train stations (and on other properties that belong to the *SBB*) to the public for rent. In some cases, space uses are already determined in advance (e.g., retail, catering). Based on currently advertised commercial spaces, it is now being examined how high the probability is that a given store will be located there. The *k kiosk* location choice prediction model (DTBG with 20 trees) was very good in terms of both specificity and sensitivity (see Section 4.2), which is why this model is used as an example. All retail spaces advertised for rent were researched from the SBB website (SBB 2024b) for which a different shop concept (e.g., catering) was not already expressly specified and which are located within train stations. Of 32 offers on the day of access (accessed on September 17,

Table 3: Performance metrics for the machine learning models

Chain	Model	Out-of-sample performance		
		Sensitivity	Specificity	ROC-AUC
Avec	BL	0.016	0.986	0.807
	DT	0.308	0.982	0.815
	DTBG*	**0.316**	**0.968**	**0.789**
	DTBG**	0.302	0.970	0.801
	DTBG***	0.294	0.971	0.805
	RF*	0.305	0.979	0.781
	RF**	0.314	0.979	0.803
	RF***	0.314	0.978	0.811
	BLGB$^{\#}$	0.144	0.991	0.828
	BLGB$^{\#\#}$	0.164	0.990	0.828
	DTGB$^{\#}$	0.090	0.998	0.876
	DTGB$^{\#\#}$	0.216	0.986	0.878
	ANN$^{\#}$	0.170	0.990	0.891
	ANN$^{\#\#}$	0.172	0.989	0.892
k kiosk	BL	0.775	0.987	0.962
	DT	0.772	1.000	0.890
	DTBG*	0.844	0.987	0.961
	DTBG**	**0.847**	**0.988**	**0.966**
	DTBG***	0.847	0.988	0.968
	RF*	0.781	0.995	0.970
	RF**	0.807	0.996	0.977
	RF***	0.804	0.996	0.978
	BLGB$^{\#}$	0.836	0.993	0.979
	BLGB$^{\#\#}$	0.824	0.992	0.977
	DTGB$^{\#}$	0.820	0.995	0.979
	DTGB$^{\#\#}$	0.824	0.994	0.982
	ANN$^{\#}$	0.821	0.990	0.983
	ANN$^{\#\#}$	0.815	0.991	0.986
Valora	BL	0.476	0.973	0.885
	DT	0.389	1.000	0.746
	DTBG*	0.627	0.951	0.878
	DTBG**	0.627	0.952	0.891
	DTBG***	0.627	0.953	0.894
	RF*	0.610	0.966	0.887
	RF**	0.622	0.966	0.897
	RF***	0.619	0.967	0.900
	BLGB$^{\#}$	**0.677**	**0.936**	**0.907**
	BLGB$^{\#\#}$	0.590	0.987	0.904
	DTGB$^{\#}$	0.525	0.985	0.932
	DTGB$^{\#\#}$	0.571	0.978	0.934
	ANN$^{\#}$	0.550	0.976	0.936
	ANN$^{\#\#}$	0.552	0.977	0.938

Notes: Models: BL = Binary Logit, DT = Decision Tree, BG = Decision Trees with Bagging, RF = Random Forest, BLGB = Binary Logit with Gradient Boosting, DTGB = Decision Trees with Gradient Boosting, ANN = Artifical Neural Network.
Flags: *, **, *** = 20, 50, or 100 trees; / $^{\#}$, $^{\#\#}$ = 100 or 200 iterations. / The best model is marked in **bold.**

2024), this applied to nine space offers. These are space offers at the following Swiss train stations (in alphabetical order): Altdorf UR, Bex, Chiasso, Genève-Eaux-Vives, Glovelier, Hedingen, Hunzenschwil, Münsingen, and St. Gallen Winkeln.

In most cases there is only one micro-location in the train station (city level), with Chiasso and Genève-Eaux-Vives being exceptions. The frequency numbers (*DTV_TJM_TGM*) of the train stations in 2023 were between 680 (Hunzenschwil) and 8,600 (Chiasso). In the case of Genève-Eaux-Vives, a commercial space was put out to tender to replace an existing food provider (*Tekoe*), which was communicated in the tender documents. This results in a change in the independent variables, namely that the value of the variable *Food_other_count* drops from 3 to 2. In the remaining cases, there is no obvious change in the business structure. In three cases, the other *Valora* format examined (*Avec*) is already located in the respective micro-location. This is important because the econometric model analysis has shown that an avoidance strategy appears to apply to these two *Valora* formats (see Section 4.1). Table 4 shows a summary of the independent variables for the train station or micro-location as well as the result of the prediction model (DTBG).

Table 4: Results of the model prediction for the opening of a *k kiosk* store

Train station		*Micro-location*			*k kiosk*
Station	DTV_TJM_TGM	level_cat	microlocation_cat	Competitors*	Prediction
Altdorf UR	2,200	0	0	2	**NO**
Bex	2,400	0	0	3**	**NO**
Chiasso	8,600	0	0	1**	**NO**
Genève-Eaux-Vives	8,400	0	2	8	**YES**
Glovelier	1,000	0	0	1	**NO**
Hedingen	2,300	0	0	2	**NO**
Hunzenschwil	680	0	0	1	**NO**
Münsingen	6,400	0	0	1**	**NO**
St Gallen Winkeln	1,500	0	0	1	**NO**

Notes: *Sum of *Avec_count*, *Migros_all_count*, *Coop_all_count*, *Disocunter_count*, *Kiosk_other_count*, *Food_other_count*, *catering_count*, and *vend_machine_count*
**Including one *Avec* or *Avec express* store (*Avec_count* > 0)

It turns out that in the nine micro-locations, a positive location decision is only predicted in one case, namely Genève-Eaux-Vives. It is unlikely that *k kiosk* stores will be opened in the remaining micro-locations, which is certainly not primarily due to insufficient demand, as the passenger frequencies in all train stations reach an acceptable level (see Section 4.1). Instead, the characteristics of the micro-locations reduce the probability of a positive result: It was already determined in the econometric analysis that *k kiosk* stores are preferred to be located in category 2 micro-locations (shopping streets, etc.), which is only the case with the commercial space on offer in Genève-Eaux-Vives. Three micro-locations are already occupied by *Avec*. The results of the model analysis suggest that there is an avoidance strategy between the two *Valora* formats *Avec* and *k kiosk*, which is why it is not surprising that *k kiosk* is unlikely to open in these locations. The fact that a food space is being abandoned in Genève-Eaux-Vives also increases the likelihood of a *k kiosk* opening. However, there is an important limitation in the results of the model forecast: In contrast to the other eight train stations, this micro-location already has a *k kiosk* store, which is not explicitly covered by the model. Although it is in principle conceivable that more than one *k kiosk* will be located at a micro-location (such as in Basel SBB, Chur, or Olten), it cannot be predicted on this basis.

5 Conclusions and limitations

The study on location choice of *Valora* convenience formats in Swiss train stations has an analytical and a predictive part. To answer the first research question, binary logit models were built for analytical purposes. The presence of *Avec* and *k kiosk* stores at the level of micro-locations within the railway stations was examined against the background of location-specific independent variables, which were derived from location theory and previous empirical work with respect to other location types. Local demand was measured by passenger frequency. The more people boarding, alighting, and transferring at a station, the more likely it is that an *Avec* or *k kiosk* store is located there. There is also a mutual avoidance strategy for both *Valora* formats. Competitors' location decisions influence the likelihood of an opening, but not to the same extent in both formats. With respect to *k kiosk*, there is no evidence of any avoidance of competition with respect to convenience supermarkets. Rather, the presence of *Coop* or *Migros*, all other things being equal, increases the probability of the presence of *k kiosk*. In the case of *Avec*, this effect is diffuse because many model parameters are not statistically significant. The most important determinants of location choice are, thus, demand and chain-specific clustering and avoidance patterns.

The positive impact of station frequency on the probability of an opening, clearly identified in all cases examined, is congruent with the statements of location theories and empirical work on both location choice and store performance, according to which market size or demand is always identified as a positive location factor. With regard to the interplay between competition and agglomeration effects, which is discussed in many theoretical and empirical contributions to retail locations, clear statements cannot be made in any case. Many approaches predict that very similar providers engage in competitor avoidance, e.g., Christaller (1933). This is obviously not the case with *k kiosk*. On the contrary, this format tends to be located where (larger) food competitors such as *Coop* or *Migros* are already located. Here, *k kiosk* may benefit from *shared business* in the sense of the *theory of cumulative attraction* (Nelson 1958) or with respect to the clustering of competitors that are imperfect substitutes (Chamberlin 1933). However, interformal competition avoidance with each other is likely for both formats, although this cannot be directly explained by location theories, as the two formats, while offering overlaps, cannot be considered completely substitutable. Rather, it is likely that a company-specific avoidance strategy is the cause.

The second research question concerned the degree to which these location decisions can be predicted in new cases. For this purpose, based on the model mentioned above, various machine learning algorithms were used to optimize the prediction ability, and the models were checked with regard to their out-of-sample accuracy. This showed that AI/ML models make a huge contribution to significantly improving the predictive ability of these models. In one case (*k kiosk*) a model was built that showed very good results in terms of both sensitivity and specificity. In the other case (*Avec*), however, it must be admitted that even the best model solution is not so good that it would be suitable for practical purposes. The suitable model was used for a forecast with real data. However, it should be noted that the selection of the best model must also follow logical considerations related to the study case. In the present case, for example, it was argued that sensitivity, i.e., predicting positive location decisions, is more important than predicting negative values. In other cases, specificity or a trade-off between the two metrics may be the decisive factor. Furthermore, it has been shown that model quality does not necessarily increase with model complexity (e.g., number of estimators) and that it is always useful to test a number of tuning parameters.

The study also faces some theoretical and methodological limitations. *Firstly*, the entire analytical and predictive model approach is based on the premise that *Valora*'s location decision is based on the evaluation of location characteristics (especially demand) and the behavior of competitors. However, it could not be taken into account whether free retail space is available at all, as there is no data on the total amount of retail space in the *SBB* train stations (i.e., including possible vacancies). It cannot therefore be clarified whether the non-presence of the examined formats at certain train stations is possibly due to the fact that opening there is not possible because of a lack of retail space. Since *Avec* is implemented on much larger sales areas than *k kiosk*, it is plausible to assume that this problem is much greater in the *Avec* model. This could in turn explain why this model performs significantly worse in terms of predictive performance than the *k kiosk* model.

Secondly, for the quantification of the demand volume, only train passenger frequencies were available, but not the actual footfall at the station (although this could potentially lead to an endogeneity problem; see Section 4.1). This is likely to underestimate the actual demand, especially in large train stations with integrated shopping streets. Furthermore, the passenger frequencies are naturally only available for the entire station, although the same frequency does not prevail in every part of the station. Passenger frequencies, therefore, do not provide a complete picture of local demand. Other variables in the model, such as the number of ticket machines, most likely partially compensate for this. Both of these limitations could be addressed in future studies. However, this would require data that is not currently (publicly) available, namely all retail spaces within stations (not just occupied ones) as well as small-scale pedestrian traffic.

Thirdly, as already mentioned, the predictive ability of the *Avec* predictive model is rather weak. At the same time, certain difficulties also appear in the analytical model in the form of some high coefficients with very high standard errors. Both problems can arise from an unfavorable combination of the explanatory variables, e.g., in terms of underspecification and/or multicollinearity. It is likely that other variables are missing here, apart from the deficit mentioned in the first point, which affects all models. For example, it could make sense to differentiate between different *Avec* subformats, e.g., *Avec* and *Avec express*, or to distinguish which *Avec* stores are open 24/7 (without service after regular opening hours) and which are not. This distinction was not made in the current study, as all *Avec* stores were treated equally. In this study, the same models were built for both convenience formats in order to be able to compare the results. However, it becomes apparent that the explanatory variables for *k kiosk* are very good but are obviously not sufficient for *Avec* and/or would have to be arranged differently in order to obtain a better model result. In principle, it's also conceivable that *Avec*'s expansion is simply less structured than *k kiosk*'s. To achieve better predictive ability for *Avec*, additional variables should be considered and/or the influence of competitors should be further differentiated, unless, as in this case, the comparison between multiple chains is the primary focus.

Fourthly, in the models, the dependent variable was coded as binary (*Valora* chain is present or not), which was calculated from the sums of the respective chains at the micro-locations. No distinction is made here as to whether one or more stores are located at the same micro-location. It is very unlikely, however, that this induces a substantial bias. In the case of *Avec*, there is no instance where more than one store is located at a micro-location. For *k kiosk*, there are only 10 micro-locations where two *k kiosk* stores are located. However, in future studies, count data models (Larsson, Oener 2014, Wieland 2017) could be used instead of binary outcome models.

References

Alonso W (1964) *Location and Land Use: Toward a General Theory of Land Rent.* Harvard University Press, Cambridge, MA. CrossRef

Altman DG, Bland JM (1994) Diagnostic tests. 1: Sensitivity and specificity. *British Medical Journal* 308: 1552. CrossRef

Aversa J, Doherty S, Hernandez T (2018) Big data analytics: The new boundaries of retail location decision making. *Papers in Applied Geography* 4: 390–408. CrossRef

Bills SJ (1998) New geographies of retailing: An investigation of developments at airports, railway stations, hospitals and service stations. PhD thesis, https://www.valora.com-/en/brands/kkiosk/

Blick (2025) Deutscher Discounter Aldi verdrängt Coop im Bahnhof Basel – Kampf um Flächen geht los. Press article, https://www.blick.ch/wirtschaft/sbb-mischen-shopping-karten-neu-deutscher-discounter-aldi-verdraengt-coop-im-bahnhof-basel-kampf-um-flaechen-geht-los-id20496591.html

Boehmke B, Greenwell B (2020) *Hands-On Machine Learning with R* (1 ed.). Taylor & Francis, New York, NY. CrossRef

Cameron AC, Trivedi PK (2005) *Microeconometrics. Methods and Applications.* Cambridge University Press, Cambridge. CrossRef

Celio Di Cellio Dias P, Forti M, Witarsa M (2018) A comparison of Gradient Boosting with Logistic Regression in Practical Cases. Working paper, https://support.sas.com/resources/papers/proceedings18/1857-2018.pdf

Chamberlin EH (1933) *The Theory of Monopolistic Competition.* Harvard University Press, Cambridge, MA

Chang HJ, Hsieh CM (2018) A new model for selecting sites for chain stores in China. *International Journal of Industrial and Systems Engineering* 28: 346–359. CrossRef

Christaller W (1933) *Die zentralen Orte in Süddeutschland: Eine ökonomisch-geographische Untersuchung über die Gesetzmäßigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen.* Gustav Fischer, Jena

Converse PD (1949) New laws of retail gravitation. *Journal of Marketing* 14: 379–384. CrossRef

DB Station&Service AG (2017) Germany's stations: Top locations for gastronomy and retail. Brochure, https://www.deutschebahn.com/resource/blob/284664/fa59e6114-fa1f1147eb699b9fe494c1f/vermietungsbroschuere˙bahnhoefe-data.pdf

Eaton BC, Lipsey RG (1982) An economic theory of central places. *The Economic Journal* 92: 56–72. CrossRef

EHI Retail Institute (2023) *Travel Retail 2023.* EHI Retail Institute

Eurostat (2024) Rail passenger transport reaches new peak in 2023. Press release, https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20241030-1

Fotheringham AS (1985) Spatial competition and agglomeration in urban modelling. *Environment and Planning A: Economy and Space* 17: 213–230. CrossRef

Fujita M, , Thisse JF (2002) *Economics of Agglomeration. Cities, Industrial Location, and Regional Growth.* Cambridge University Press, Cambridge. CrossRef

Ge D, Hu L, Jiang B, Su G, Wu X (2019) Intelligent site selection for bricks-and-mortar stores. *Modern Supply Chain Research and Applications* 1: 88–102. CrossRef

Ghosh A (1986) The value of a mall and other insights from a revised central place model. *Journal of Retailing* 62: 79–97

Greene WJ (2012) *Econometric Analysis* (7 ed.). Pearson

Hlavac M (2022) stargazer: Well-formatted regression and summary statistics tables. R package version 5.2.3. Software, https://CRAN.R-project.org/package=stargazer, Bratislava, Slovakia

Hotelling H (1929) Stability in competition. *The Economic Journal* 39: 41–57. CrossRef

Huff DL (1962) *Determination of Intra-Urban Retail Trade Areas.* University of California

Joseph L, Kuby M (2013) Regionalism in US retailing. *Applied Geography* 37: 150–159. CrossRef

Joseph L, Kuby M (2015) Modeling retail chain expansion and maturity through wave analysis: Theory and application to Walmart and Target. *International Journal of Applied Geospatial Research* 6: 1–26. CrossRef

Joseph L, Kuby M (2016) The location types of US retailers. *International Journal of Applied Geospatial Research* 7: 1–22. CrossRef

Krider R, Putler D (2013, 04) Which birds of a feather flock together? Clustering and avoidance patterns of similar retail outlets. *Geographical Analysis* 45: 123–149. CrossRef

Kuhn M (2008) Building predictive models in R using the caret package. *Journal of Statistical Software* 28: 1–26. CrossRef

Larsson JP, Oener O (2014) Location and co-location in retail: A probabilistic approach using geo-coded data for metropolitan retail markets. *The Annals of Regional Science* 52: 385–408. CrossRef

Liaw A, Wiener M (2002) Classification and regression by randomforest. *R News* 2: 18–22

Lu J, Zheng X, Nervino E, Li Y, Xu Z, Xu Y (2024) Retail store location screening: A machine learning-based approach. *Journal of Retailing and Consumer Services* 77: 103620. CrossRef

Müller-Hagedorn L (1991) Moderne Verfahren zur Ermittlung der Bedeutung einzelner Standortfaktoren. In: DHI (ed), *Standortpolitik des Einzelhandels*. Köln, 100–105

Nelson P (1970) Information and consumer behavior. *Journal of Political Economy* 78: 311–329. CrossRef

Nelson RL (1958) *The Selection of Retail Locations*. F.W. Dodge, West Palm Beach

Neue Zürcher Zeitung (2024) Discounter unerwünscht? Die SBB lassen Aldi und Lidl abblitzen. Press article, https://www.nzz.ch/wirtschaft/discounter-unerwuenscht-lidl-und-aldi-bewerben-sich-bei-den-sbb-um-ladenflaechen-aber-blitzen-ab-ld.1803620

Nilsson IM, Smirnov OA (2016) Measuring the effect of transportation infrastructure on retail firm co-location patterns. *Journal of Transport Geography* 51: 110–118. CrossRef

OEBB (2024) Immobilien-Angebote. Website, https://immobilien.oebb.at/de/angebote

Office of Rail and Road (2024) Railway station catering market. Final Report. https://www.orr.gov.uk/sites/default/files/2024-06/railway-station-catering-market-study-final-report-june-2024_0.pdf

Omar ED, Mat H, Zafirah Abd Karim A, Sanaudi R, Ibrahim FH, Azahadi Omar M, Zulfadli Hafiz Ismail M, Jayaraj VJ, Leong Goh B (2024) Comparative analysis of logistic regression, gradient boosted trees, svm, and random forest algorithms for prediction of acute kidney injury requiring dialysis after cardiac surgery. *International Journal of Nephrology and Renovascular Disease* 17: 197–204. CrossRef

Orhun AY (2013) Spatial differentiation in the supermarket industry: The role of common information. *Quantitative Marketing and Economics* 11: 3–37. CrossRef

R Core Team (2024) R: A Language and Environment for Statistical Computing. Software, https://www.R-project.org/, Vienna, Austria

Radio Frequence Jura (2024) La gare de Delémont se sépare de son Coop Pronto. Press article, https://www.rfj.ch/rfj/Actualite/Region/20240127-La-gare-de-Delemont-se-separe-de-son-Coop-Pronto.html

Rao F, Pafka E (2021) Shopping morphologies of urban transit station areas: A comparative study of central city station catchments in Toronto, San Francisco, and Melbourne. *Journal of Transport Geography* 96: 103156. CrossRef

Reigadinha T, Godinho P, Dias J (2017) Portuguese food retailers – Exploring three classic theories of retail location. *Journal of Retailing and Consumer Services* 34: 102–116. CrossRef

Reilly WJ (1931) *The Law of Retail Gravitation*. Knickerbocker Press

Reynolds J, Wood S (2010) Location decision making in retail firms: Evolution and challenge. *International Journal of Retail & Distribution Management* 38: 828–845. CrossRef

Rice MD, Ostrander A, Tiwari C (2016) Decoding the development strategy of a major retailer: Wal-Mart's expansion in the United States. *The Professional Geographer* 68: 640–649. CrossRef

SBB (2024a) Ein- und Aussteigende an Bahnhöfen. Dataset, https://data.sbb.ch/explore-/dataset/passagierfrequenz

SBB (2024b) Freie Retailflächen. Website, https://sbb-immobilien.ch/mieten/retail/

SBB (2024c) SBB Company. Website, https://company.sbb.ch/en/home.html

SBB (2024d) Öffnungszeiten Shops. Dataset, https://data.sbb.ch/explore/dataset/offnungszeiten-shops/

Seong EY, Lim Y, Choi CG (2022) Why are convenience stores clustered? The reasons behind the clustering of similar shops and the effect of increased competition. *Environment and Planning B: Urban Analytics and City Science* 49: 834–846. CrossRef

Shahidi F, Rennert-May E, D'Souza AG, Crocker A, Faris P, Leal J (2023) Machine learning risk estimation and prediction of death in continuing care facilities using administrative data. *Scientific Reports* 13: 17708. CrossRef

Simkin L (1989) SLAM: Store location assessment model - Theory and practice. *Omega-international Journal of Management Science* 17: 53–58. CrossRef

SNCF Gares & Connexions (2024) The station, new town centre venue. Website, https:-//www.garesetconnexions.sncf/en/retail/retail-commercial-activity

Stanley TJ, Sewall MA (1976) Image inputs to a probabilistic model: Predicting retail potential. *Journal of Marketing* 40: 48–53. CrossRef

Tabuchi T, Thisse JF (2011) A new economic geography model of central places. *Journal of Urban Economics* 69: 240–252. CrossRef

Taylor RD (1978) Retail site selection using multiple regression analysis. Phd thesis, https://sbb-immobilien.ch/mieten/retail/

Themido IH, Quintino A, Leitão J (1998) Modelling the retail sales of gasoline in a Portuguese metropolitan area. *International Transactions in Operational Research* 5: 89–102. CrossRef

Ting CY, Jie MY (2022) Location profiling for retail-site recommendation using machine learning approach. In: Haw SC, Muthu KS (eds), *Proceedings of the International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022)*, Volume 10 of *Atlantis Highlights in Computer Sciences*. Atlantis Press, 85–116. CrossRef

Trevethan R (2017) Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health* 5. CrossRef

Turhan G, Akalın M, Zehir C (2013) Literature review on selection criteria of store location based on performance measures. *Procedia - Social and Behavioral Sciences* 99: 391–402. CrossRef

Valora Holding AG (2019) Andermatt: Zweigrösster neuer avec Store der Schweiz feierlich eingeweiht. Press release, https://www.avec.ch/media/standorte/eroeffnungen/-20191220_val_mm_avec_andermatt.pdf

Valora Holding AG (2024a) avec - "Handmade with Love". Website, https://www.valora.com/en/brands/avec/

Valora Holding AG (2024b) Immobilien. Website, https://www.valora.com/de/contact/-immobilien/

Valora Holding AG (2024c) k kiosk - “Treat yourself”. Website, https://www.valora.com-/en/brands/kkiosk/

Valora Holding AG (2024d) Our brands. Website, https://www.valora.com/en/brands/

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S* (Fourth ed.). Springer, New York. CrossRef

Wang L, Fan H, Wang Y (2018) Site selection of retail shops based on spatial accessibility and hybrid BP neural network. *ISPRS International Journal of Geo-Information* 7. CrossRef

Weber B (1979) *Eine statistische Analyse der Abhängigkeiten des Kundenaufkommens von Standorteinflüssen bei Einzelhandelsgeschäften. Dargestellt an ausgewählten Apotheken der Stadt Münster*, Volume 45 of *Schriftenreihe wirtschaftswissenschaftliche Forschung und Entwicklung*. Florentz

Wieland T (2017) Versorgungsstrukturen und Tragfähigkeit von Gesundheitseinrichtungen aus einer standortökonomischen Perspektive. In: Harteisen U, Dittrich C, Reeh T, Eigner-Thiel S (eds), *Land und Stadt - Lebenswelten und planerische Praxis*, Volume 121 of *Göttinger geographische Abhandlungen*. Goltze, 85–116

Wieland T (2018) Standorterfolg in Zeiten des Onlinehandels - Aufbau, Ergebnisse und planungsbezogene Implikationen einer modellgestützten Standortanalyse für die Elektrofachmärkte in der Region Mittlerer Oberrhein. *Berichte. Geographie und Landeskunde* 92: 5–26. https://www.geographische-regionalforschung.de/download/-1254/bd-92-heft-1/1481/bgl_bd_92_heft_1_2018_01_wieland.pdf

Wieland T (2023) Spatial shopping behavior during the Corona pandemic: Insights from a micro-econometric store choice model for consumer electronics and furniture retailing in Germany. *Journal of Geographical Systems* 25: 291–326. CrossRef

Wolinsky A (1983) Retail trade concentration due to consumers’ imperfect information. *The Bell Journal of Economics* 14: 275–282. CrossRef

Zhou L, Wang S, Li H (2024) Store network expansion in the era of online consumption: Evidence from the suning appliance retail chain in China. *Applied Geography* 165: 103225. CrossRef

Zhou Q, Huang K, Huang D (2015) Forecasting sales using store, promotion, and competitor data. Report, https://jmcauley.ucsd.edu/cse255/projects/fa15/022.pdf

Zhu T, Singh V (2009) Spatial competition with endogenous location choices: An application to discount retailing. *Quantitative Marketing and Economics* 7: 1–35. CrossRef

REGION

The Journal of ERSA
Powered by WU

Volume 12, Number 2, 2025, 23–50
DOI: 10.18335/region.v12i2.566

journal homepage: region.ersa.org
ISSN: 2409-5370

Sequence Analysis of Neighborhood Racial and Ethnic Changes: The Case of New York City 1980-2020

Elizabeth Delmelle[1], **Eric Delmelle**[2]

[1] University of Pennsylvania
[2] Lehigh University and Vrije University Brussels

Received: October 2, 2024/Accepted: October 12, 2025

Abstract. This paper demonstrates the application of sequence analysis to develop a typology of racial and ethnic trajectories in New York City neighborhoods from 1980 to 2020 using a reproducible R workflow. Our workflow begins with using an unsupervised classification method, k-means, at each decennial cross-section to derive 6 classes describing the racial and ethnic makeup of neighborhoods during the study period. These classes include four that depict a majority of Black, White, Hispanic, and Asian residents, and two mixed-race classes, Black and Hispanic, and a White majority with a mixture of other races. We then develop a sequence of classes for each census tract over the 5 decennial time stamps. Finally, we derive a longitudinal typology describing the predominant pathways of change using sequence analysis. This resulted in 14 distinct pathways including transitions to Hispanic and Asian majorities emerging from historically White or Black neighborhoods. The findings underscore the gradual nature of neighborhood racial transformations. Our approach is reproducible for researchers wanting to explore and visualize multidimensional neighborhood dynamics.

1 Introduction

Tracking and understanding neighborhood changes has been a central topic of urban studies and a fundamental concern for planning practitioners (Galster 2001, Landis 2016, Chapple, Zuk 2016). Neighborhood change can be comprehended according to various dimensions, from housing stock or built environment changes to residents' demographic and socioeconomic composition (Delmelle 2022). Thus, the study of neighborhood change involves analyzing multiple attribute dimensions through time for spatially situated units. Recent scholarship has progressed in the analytical strategies used to study neighborhood trajectories, introducing new methods for visualizing and mapping longitudinal pathways of change for multiple dimensions (Delmelle 2022).

In this article, we demonstrate one such technique, sequence analysis, in an illustrative, reproducible tutorial analyzing neighborhood change. We demonstrate the method using a case study of racial and ethnic changes in New York City census tracts from 1980-2020. While our focus is methodological, the empirical case study highlights the utility of sequence analysis in visualizing, detecting, and exploring longitudinal patterns of neighborhood changes.

Since the 1980s, the United States' demographic profile has become increasingly diverse in the past several decades, driven by sustained immigration and by the growing share of births to non-White populations, influencing demographic shifts in the overall population structure (Frey 2022). Analyses following the release of the 2010 decennial census showed that increasing national diversity resulted in differing neighborhood trajectories, largely contingent on the broader metropolitan context (Terbeck 2023, Wright et al. 2014).

Much of the empirical scholarship on neighborhood racial and ethnic changes has developed indices to categorize the makeup or diversity of racial and ethnic groups and then explored changes in a neighborhood's diversity categorization over time. For example, a neighborhood might transition from 'low diverse' to 'moderately diverse' from one decade to the next (Farrell, Lee 2011, Wright et al. 2014). Alternately, another set of techniques aims to describe the longitudinal sequences or trajectories that the racial and ethnic groups in a neighborhood have followed. Three methods have been adopted in the literature for this purpose: statistical curve fitting, time-series clustering, and sequence analysis. With curve-fitting models like growth mixture models or latent growth models, mathematical functions fit each racial and ethnic group under study that summarize the predominant trends (Zwiers et al. 2018, Hipp, Kim 2023). Time series clustering is an unsupervised classification technique aimed at clustering continuous longitudinal data (Delmelle et al. 2025). Finally,in sequence analysis, neighborhoods are first grouped into similar categorical clusters at each time stamp. Then, each a sequence of neighborhood categories is constructed over time. Finally, the sequences of these clusters are grouped using a sequence alignment technique (Delmelle 2016, González-Leonardo et al. 2023).

While all three approaches - curve fitting, time-series clustering, and sequence analysis aim to characterize change over time, they differ in what types of data and research questions they are best suited for. Curve fitting methods are well-suited for summarizing trends in continuous outcomes, especially when the trajectory is expected to follow a smooth or parametric form, which must be specified *a priori*. Time-series clustering also retains the continuous nature of longitudinal data, but must be performed on single variables at a time. Cross-sectional classes of trajectories can be constructed to form multivariate typologies (Delmelle et al. 2025), however, as the number of variables increases, this becomes an arduous workflow.

Sequence analysis is oriented towards categorical trajectories that can be constructed using multiple input variables, as demonstrated in this article. The method is particularly useful in visualizing and analyzing the timing, order, and transitions between qualitatively distinct neighborhood states rather than in the smoothness of changes.

In this article, we explore trajectories of neighborhood racial and ethnic changes in the largest and one of the most diverse cities in the United States, New York City, using longitudinal census data up to the latest 2020 decennial release. This notebook showcases a workflow that introduces sequence analysis to the study of multidimensional neighborhood changes. We begin by processing the raw longitudinal census data, then performing a k-means classification, and finally classify and map sequences clusters. Our case study illustrates the gradual progression that neighborhoods follow in when undergoing racial and ethnic transformations. We observe an overall decline in the share of neighborhoods categorized by a large White majority population, in exchange for increasing diversity, especially Hispanic and Asian populations.

2 Computational environment

The main libraries used in this paper include `dplyr` for processing tabular census data, `sf` for mapping the resulting clusters, `cluster` for performing the `k-means` cluster analysis. Finally, to perform the sequence analysis, we use `TraMineR` (Gabadinho et al. 2011). This is a popular and continually updated R package for performing sequence analysis for a host of social science applications, including analyses of neighborhood change (Delmelle 2016, 2017, Patias et al. 2020).

[1]:

```r
# Set CRAN mirror
options(repos = c(CRAN = "https://cloud.r-project.org/"))

# For data management
if (!require('knitr')) install.packages('knitr'); library('knitr')
if (!require('rmarkdown')) install.packages('rmarkdown'); library('rmarkdown')
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
```

```
#for reading in data
if (!require('here')) install.packages('here'); library('here')

#for reading in census data
if (!require('tidycensus')) install.packages('tidycensus'); library('tidycensus')
if (!require('tigris')) install.packages('tigris'); library('tigris')
options(tigris_class = "sf") # returns data in sf format

#for data table formatting
if (!require('knitr')) install.packages('knitr'); library('knitr')

#for data table pivoting
if (!require('tidyverse')) install.packages('tidyverse'); library('tidyverse')

#mapping packages
if (!require('sp')) install.packages('sp'); library('sp')
if (!require('sf')) install.packages('sf'); library('sf')

#For data visualization
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')

#for facet plots
if (!require('gridExtra')) install.packages('gridExtra'); library('gridExtra')

#For sequence clustering
if (!require('TraMineR')) install.packages('TraMineR'); library('TraMineR')

# For k-means and hierarchical cluster analysis
if (!require('cluster')) install.packages('cluster'); library('cluster')

#For visualizing k-means outputs
if (!require('factoextra')) install.packages('factoextra'); library('factoextra')

# For creating heat map to describe k-means cluster results
if (!require('pheatmap')) install.packages('pheatmap'); library('pheatmap')

# For creating heat map to describe k-means cluster results
if (!require('RColorBrewer')) install.packages('RColorBrewer'); library('RColorBrewer')

# Other packages (cowplot:inset maps; extrafont for additional sans serif fonts)
if (!require('cowplot')) install.packages('cowplot'); library('cowplot')
if (!require('patchwork')) install.packages('patchwork'); library('patchwork')
if (!require('extrafont')) install.packages('extrafont'); library('extrafont')
loadfonts(device = "win")
if (!require('ggspatial')) install.packages('ggspatial'); library('ggspatial')
loadfonts(device = "win")
```

3 Data

We use decennial census tract data to examine neighborhood racial and ethnic changes. Census tracts serve as imperfect, yet well-used neighborhood proxies. Census tract boundaries change over time, further complicating the study of population dynamics within these boundaries. There are several sources of data that have been harmonized using interpolation techniques to a consistent set of boundaries over time. We use the Longitudinal Tract Database (LTDB) which uses areal and population interpolation techniques alongside ancillary data on water cover to derive estimates (Logan et al. 2014). Analyses of the errors produced by three popular longitudinal data providers suggest that LTDB performs similarly to the dataset produced by the National Historic Geographic Information System (NHGIS) and both perform better than the Neighborhood Change Database which relies solely on areal interpolation without the inclusion of ancillary data (Logan et al. 2014). Therefore, for this type of analysis either the LTDB or NHGIS would be suitable dataset for this analysis.

We obtained the full count decennial data from LTDB from 1980-2020 from the Diversity and Disparties project at Brown University. The census variables have been interpolated to 2010 tract boundaries. Because the coding of census race and ethnicity changes over time, we opted to begin in 1980 as 1970, the earliest dataset available, did not record a count of Latino or Hispanic residents. The raw data contains all census tracts throughout the United States. For visualization purposes, we also import a shapefile of 2010 census tract boundaries using the `tidycensus` package and setting the geometry to `true`.

[2]:
```
setwd(here())  #current working directory

#csv tables for longitudinal data
census20<- read.csv("data/ltdb_std_2020_fullcount.csv")
census10<- read.csv("data/LTDB_Std_2010_fullcount.csv")
census00<- read.csv("data/LTDB_Std_2000_fullcount.csv")
census90<- read.csv("data/LTDB_Std_1990_fullcount.csv")
census80<- read.csv("data/LTDB_Std_1980_fullcount.csv")

#geometry data
#filter for NYC counties (5 boroughs)
nyc_counties <- c("Kings", "Queens", "New York", "Richmond", "Bronx")
tract <- get_decennial(geography = "tract",
                       variables = "P001001",
                       year = 2010,
                       state = "NY",
                       county = nyc_counties,
                       geometry = TRUE,
                       progress = FALSE)
```

We next calculate the share of `White`, `Black`, `Hispanic`, and `Asian` residents in each tract for each decade from the raw count using the total population as the denominator. We filter out tracts with no population and select only the relevant columns to create our data frame. We then join all columns from the five decennial data frames into one data frame called `census_all` and finally select only census tracts from the five counties that comprise New York City's five boroughs: Bronx County, Kings County (Brooklyn), New York County (Manhattan), Queens County, Richmond County (Staten Island).

[3]:
```
census80 <- census80 %>% filter (POP80 >0)
census80$perwhite80 <- census80$NHWHT80/census80$POP80
census80$perblack80 <- census80$NHBLK80/census80$POP80
census80$perhisp80 <- census80$HISP80/census80$POP80
census80$perasian80 <- census80$ASIAN80/census80$POP80
census80<- census80 %>% select(c("TRTID10","perwhite80", "perblack80", "perhisp80",
                                 "perasian80"))

census90 <- census90 %>% filter (POP90 >0)
census90$perwhite90 <- census90$NHWHT90/census90$POP90
census90$perblack90 <- census90$NHBLK90/census90$POP90
census90$perhisp90 <- census90$HISP90/census90$POP90
census90$perasian90 <- census90$ASIAN90/census90$POP90
census90<- census90 %>% select(c("TRTID10","state","county","perwhite90", "perblack90",
                                 "perhisp90", "perasian90"))

census00 <- census00 %>% filter (POP00 >0)
census00$perwhite00 <- census00$NHWHT00/census00$POP00
census00$perblack00 <- census00$NHBLK00/census00$POP00
census00$perhisp00 <- census00$HISP00/census00$POP00
census00$perasian00 <- census00$ASIAN00/census00$POP00
census00<- census00 %>% select(c("TRTID10","perwhite00", "perblack00", "perhisp00",
                                 "perasian00"))

census10 <- census10 %>% rename("TRTID10" = "tractid")
census10$perwhite10 <- census10$nhwht10/census10$pop10
census10$perblack10 <- census10$nhblk10/census10$pop10
census10$perhisp10 <- census10$hisp10/census10$pop10
census10$perasian10 <- census10$asian10/census10$pop10
census10<- census10 %>% select(c("TRTID10","perwhite10", "perblack10", "perhisp10",
                                 "perasian10"))

census20 <- census20 %>% rename("TRTID10" = "TRTID2010")
census20$perwhite20 <- census20$nhwt20/census20$pop20
census20$perblack20 <- census20$nhblk20/census20$pop20
census20$perhisp20 <- census20$hisp20/census20$pop20
```

```
census20$perasian20 <- census20$asian20/census20$pop20
census20<- census20 %>% select(c("TRTID10","perwhite20", "perblack20", "perhisp20",
                                 "perasian20"))

#join all data frames from each decade
census_all<- census90 %>%
  left_join(census00) %>%
  left_join(., census10) %>%
  left_join(., census20)%>%
  left_join(., census80)

#Select NYC Counties. These include Bronx County, Kings County (Brooklyn),
#New York County (Manhattan), Queens County, Richmond County (Staten Island)

census_select <- census_all %>% filter((state == "NY" & county == "Bronx County")|
                              (state == "NY" & county == "Kings County")|
                              (state == "NY" & county == "New York County")|
                              (state == "NY" & county == "Queens County")|
                              (state == "NY" & county == "Richmond County"))

##remove NA values and state and county columns
census_nyc <- na.omit(census_select)%>% select(-state, -county)

#Select only the tractID column from the shapefile. Rename the field for ease of joining
#and convert to double to match the csv data.
tract<- tract %>% select("GEOID")
tract<- rename(tract, TRTID10 = GEOID)
tract$TRTID10<- as.double(tract$TRTID10)
```

[4]:

```
# Set tigris options to return the data in sf format (spatial data frame)
options(tigris_class = "sf")

# Download county boundaries for New York state
ny_counties <- counties(state = "NY", cb = TRUE,
                         progress = FALSE)

# Filter for only the New York City counties
nyc_counties <- ny_counties %>%
  filter(NAME %in% c("Bronx", "Kings", "New York", "Queens", "Richmond"))

# View the structure of the data
print(nyc_counties)
```

[4]:

```
Simple feature collection with 5 features and 12 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -74.25563 ymin: 40.4961 xmax: -73.70036 ymax: 40.91771
Geodetic CRS:  NAD83
  STATEFP COUNTYFP COUNTYNS        GEOIDFQ GEOID     NAME        NAMELSAD
1      36      005 00974101 0500000US36005 36005    Bronx    Bronx County
2      36      047 00974122 0500000US36047 36047    Kings    Kings County
3      36      061 00974129 0500000US36061 36061 New York New York County
4      36      081 00974139 0500000US36081 36081   Queens   Queens County
5      36      085 00974141 0500000US36085 36085 Richmond Richmond County
  STUSPS STATE_NAME LSAD     ALAND    AWATER                       geometry
1     NY   New York   06 109235672  39353304 MULTIPOLYGON (((-73.77242 4...
2     NY   New York   06 179684481  71158757 MULTIPOLYGON (((-74.04171 4...
3     NY   New York   06  58683879  29010416 MULTIPOLYGON (((-74.00641 4...
4     NY   New York   06 281594051 188444349 MULTIPOLYGON (((-73.96262 4...
5     NY   New York   06 148982679 117441532 MULTIPOLYGON (((-74.16154 4...
```

4 Basic conceptual intuition

4.1 Categorizing Neighborhoods Using k*-means*

Sequence analysis requires categorical input states. Given that neighborhood demographic and socioeconomic data is very often continuous, the first step is to transform the data into discrete categories for each year of the analysis. In this example, we opt to use a data-driven, unsupervised classification approach for this stage by applying k-means to cluster the racial and ethnic composition of each census tract in each decade into one of several mutually exclusive categories. This step serves as a preprocessing stage, not an analysis of changes itself.

As an alternative, for the case of racial and ethnic compositions, for example, prespecified thresholds (e.g., >50% of a population group) could be used to define categories like “majority Black” or “majority Hispanic”. This approach works well in situations where clear theoretical cutoffs exist. However, other applications involving socio-economic indicators or multivariate neighborhood characteristics may not lend themselves to clear analyst-defined cutoffs, making a data-driven approach preferable. Since the purpose of this article is illustrative, we proceed with an overview of implementing k-means to construct the initial discrete classes of neighborhoods from the original racial and ethnic makeup data. However, if the analyst already has theoretically grounded thresholds, they can bypass this step and proceed directly to the sequence analysis.

In the context of neighborhood racial and ethnic classification, Reibel, Regelson (2011) introduced the idea of using an unsupervised classification approach for studying neighborhood racial change as an alternative to the use of neighborhood diversity indices. The objective of the k-means algorithm is to group observations in such a way that maximizes the similarity of observations within groups or clusters while maximizing the dissimilarity between each cluster. In other words, the goal is to group neighborhoods so that those most similar to each other according to their racial and ethnic makeup are assigned to the same cluster and the clusters themselves are distinct from one another in terms of their makeup.

With the k-means algorithm, the number of clusters, k must be determined *a priori*. To make this determination, it is customary to evaluate multiple solutions using various fit statistics in conjunction with domain and geographic knowledge of the data (Delmelle 2015). It is also commonly recommended that input variables first be normalized to avoid placing unequal emphasis on variables that may be on different measurement scales. However, in our case study, all of the racial and ethnic variables represent percentages of the population and so this step is not performed in the case study.

Our ultimate goal is to understand the major pathways of neighborhood change and so each neighborhood will be classified five times for 1980, 1990, 2000, 2010, and 2020 to establish its longitudinal sequence. To ensure that the clusters are temporally stable, we will perform the clustering for all years at once. New neighborhood typologies may emerge over time with this approach. In that instance, only tracts from the later years would be assigned to the new cluster.

4.2 Identifying pathways of change using sequence analysis

Like k-means, sequence analysis is an unsupervised classification technique but instead of grouping observations based on cross-sectional similarity, it clusters them based on the similarity of longitudinal categorical sequences. A central component of sequence analysis is defining how similar or dissimilar two sequences are. There are multiple methods to compute sequence dissimilarity. Studer, Ritschard (2016) provides a comprehensive overview of techniques for determining sequences dissimilarity for social science applications.

In this case study, our goal is to group neighborhoods that follow similar racial and ethnic change trajectories, specifically in which order of categories traverse through. This gives us an overview of the various pathways of change neighborhoods may take. For example sequences that progress from a majority White composition to Mixed Race and eventually to majority Hispanic may represent one frequent change pathway. Because our sequences are measured at fixed intervals (every decade from 1980 to 2020), and each neighborhood has the same number of observations, we are less concerned with irregular timing or duration, which are often central considerations in life-course studies.

To prioritize this kind of ordered similarity, we use OMstrans, a variation of the popular Optimal Matching (OM) algorithm (Delmelle 2017). OM treats sequences as strings and calculates the ‘edit distance’, or minimal cost to transform one sequence into another using insertions, deletions, and substitutions. In OMstrans, the algorithm operations not on the raw states, but on the transitions between states by merging eaech state in a sequence with its predecessor. This approach emphasizes how one state leads into another and helps preserve the sequencing of change.

For example, one neighborhood may have a sequence of *White*, *White*, *Mixed Race*, *Hispanic* while another neighborhood might follow the sequence *White*, *White*, *Mixed Race*, *Mixed Race*. Between these two strings, there is one entry that differs: the final state. To transform one to another, we could substitute *Hispanic* for *Mixed Race* in the second sequence, and the dissimilarity of the two sequences would be equal to the cost of that substitution.

In the OMstrans variant, distances between sequences of transitions are computed. This means that each state is merged with its previous state to create a subsequence. In the previous illustrative example, the first sequence becomes (*White-White*, *White-MixedRace*, *MixedRace-Hispanic*) and the OM cost evaluation is then applied to these subsequences.

The OMstrans approach allows us to incorporate empirical transition probabilities into the substitution cost matrix. In this way, frequent transitions (e.g., *White* to *Mixed*) are penalized less than rare ones. Insertions and deletions can be discoraged by assigning a higher cost than the empirical transitions so that disruptions to the temporal alignment of sequences is discouraged. The balance between preserving state similarity and order is further governed by the parameter w, the origin-transition parameter. When $w = 1$, OMstrans approximates the traditional OM algorithm. A lower value places greater emphasis on the ordering or sequencing of events than on the specific states themselves.

5 Application

5.1 Study Area

Our study area consists of census tracts within the five Boroughs of New York (see Figure 1 for an overview map).

[5]:

```
# Filter for only the New York City counties
nyc_counties <- ny_counties %>%
  filter(NAME %in% c("Bronx", "Kings", "New York", "Queens", "Richmond"))

# Modify labels for specific counties
nyc_counties$label_main <- ifelse(nyc_counties$NAME == "New York", "New York",
                                  ifelse(nyc_counties$NAME == "Kings", "Kings",
                                         ifelse(nyc_counties$NAME == "Richmond",
                                                "Richmond", nyc_counties$NAME)))

nyc_counties$label_paren <- ifelse(nyc_counties$NAME == "New York", "(Manhattan)",
                                   ifelse(nyc_counties$NAME == "Kings", "(Brooklyn)",
                                          ifelse(nyc_counties$NAME == "Richmond",
                                                 "(Staten Island)", "")))

# Calculate centroids for each county to get coordinates for labels
centroids <- st_centroid(nyc_counties)
coords <- st_coordinates(centroids)

# Add the coordinates to the nyc_counties data frame
nyc_counties <- nyc_counties %>%
  mutate(X = coords[, 1],
         Y = coords[, 2])

# Adjust coordinates manually for specific counties
nyc_counties <- nyc_counties %>%
  mutate(
    X = ifelse(NAME == "New York", X - 0.11,               # Move "New York" left,
               ifelse(NAME == "Richmond", X + 0.012, X)), # Richmond right
    Y = ifelse(NAME == "New York", Y + 0.005,
               ifelse(NAME == "Richmond", Y + 0.03, Y)) # Adjust Richmond label higher
  )

#you can mask water area if you would like.
#nyc_counties <- nyc_counties %>%  erase_water(area_threshold = 0.9)

# Create the large map for the state without a scale bar
zoomed_map <- ggplot() +
  geom_sf(data = ny_counties, fill = "gray85", color = "white", size = 1) +
            # New York State outline
  geom_sf(data = nyc_counties, fill = "#85B0A9", color = "white", size = 3) +
            # NYC counties in teal
  geom_segment(aes(x = -73.98, y = 40.775, xend = -74.03, yend = 40.775),
            color = "gray25", size = 0.5) +
            # Line from New York label to the boundary
  geom_text(data = nyc_counties, aes(x = X, y = Y, label = label_main),
            color = "black", size = 4, fontface = "bold") +
            # Add main county names
  geom_text(data = nyc_counties, aes(x = X, y = Y - 0.02, label = label_paren),
            color = "gray25", size = 3.75) +
            # Add parentheses labels in lighter color
  geom_text(aes(x = -73.6, y = 40.725, label = "Nassau"),
            color = "gray55", size = 4, fontface = "bold") +
            # Corrected Nassau label position
  geom_text(aes(x = -73.8, y = 40.98, label = "Westchester"),
            color = "gray55", size = 4, fontface = "bold") +
            # Corrected Westchester label
  geom_text(aes(x = -73.6, y = 40.705, label = "(Long Island)"),
            color = "gray70", size = 3.5, fontface = "bold") +
            # Add Long Island label below Nassau
  coord_sf(xlim = c(-74.25, -73.5), ylim = c(40.4, 41.05), expand = FALSE) +
            # Crop to focus on NYC
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(), # Remove major grid lines
    panel.grid.minor = element_blank(), # Remove minor grid lines
    axis.title = element_blank(), # Remove axis titles
    axis.text = element_blank(), # Remove axis text (labels)
    legend.position = "none", # Remove the legend
    plot.margin = margin(0, 0, 0, 0) # Remove margins for a tighter crop
  )

# Create the large map for NY and surrounding counties with text annotation
inset_map <- ggplot() +
  geom_sf(data = ny_counties, fill = "gray90", color = NA, size = 1) +
            # All counties in gray, including Nassau and Westchester
  geom_sf(data = nyc_counties, fill = "#569289", color = "#569289", size = 5) +
            # NYC counties in darker teal
  annotate("text", x = -76.2, y = 42.8, label = "New York State",
           size = 5.25, color = "gray49", angle = 0,
           alpha = 0.7) + # Add text label
  theme_minimal() +
  theme(
    panel.grid.major = element_blank(), # Remove major grid lines
    panel.grid.minor = element_blank(), # Remove minor grid lines
    plot.title = element_text(hjust = 0), # Align title to the left
    plot.subtitle = element_text(hjust = 0), # Align subtitle to the left
    axis.title = element_blank(), # Remove axis titles
    axis.text = element_blank(), # Remove axis text (labels)
    plot.margin = margin(0, 0, 0, 0) # Tighten margins
  )

# Combine the maps, placing the inset map within the zoomed-in map
final_plot <- ggdraw() +
  draw_plot(zoomed_map) +
  draw_plot(inset_map, x = 0.07, y = 0.65, width = 0.3, height = 0.3)
            # Adjust position and size of the inset

# Display the combined plot
print(final_plot)
```

[5]:

```
Output in Figure 1
```

To begin our case study, we start by preparing the data for the *k*-means clustering. This involves pivoting the data frame so that each census tract is represented with five distinct rows, once for each decennial value. To do so, we pivot from a wide to a long format and select out just the four race and ethnicity values to be used in the clustering.

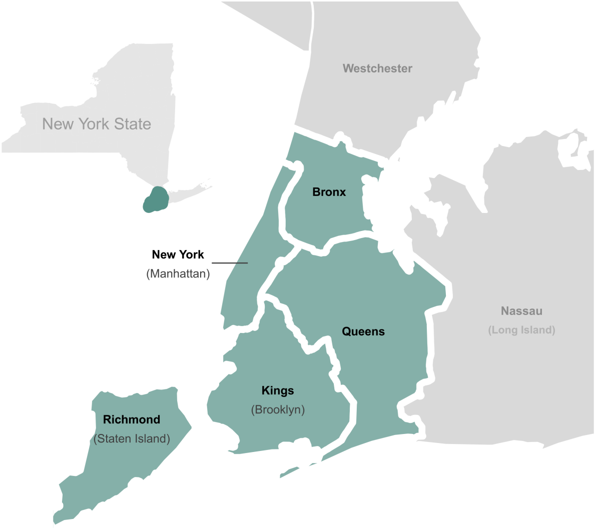

Figure 1: Overview map of New York City's five boroughs and surrounding counties

[6]:

```r
# Convert the data frame from wide to long format
census_long <- census_nyc %>%
  pivot_longer(cols = starts_with("per"),
               names_to = c(".value", "year"),
               names_pattern = "per(\\w+)(\\d{2})") %>%
  mutate(year = case_when(
    year == "80" ~ 1980,
    year == "90" ~ 1990,
    year == "00" ~ 2000,
    year == "10" ~ 2010,
    year == "20" ~ 2020,
    TRUE ~ as.integer(year)
  ))

data_for_clustering <- census_long %>%
  select(white, black, hisp, asian)
```

As explained above, the k-means clustering procedure requires that the number of clusters, k be specified *a priori*. There are fit statistics that attempt to determine the optimal number of clusters, considering the similarity of observations within each cluster and the distinctiveness of the clusters from each other. However, these mathematically-derived solutions are devoid of any contextual or theoretical understanding of the problem under study. Therefore, the selection of k often becomes more akin to art than a science (Von Luxburg et al. 2012), considering the objective of the study. We begin with two common data-driven approaches that explore multiple clustering solutions for different k values and then examines the *within sum of squares* (WSS) and the *average silhouette score* for each solution. The WSS assesses how compact a clustering solution is, or how homogeneous the observations assigned to each cluster are, and the average silhouette score measures how well separated each cluster is from each other. Our objective is to derive a typology of neighborhoods according to their racial and ethnic makeup, for four groups: percent White, Black, Hispanic, and Asian. Figures 2 and 3 show the results of the clustering analysis; the Elbow method plot (Figure 2) suggests the optimal number of clusters, while the Silhouette method plot (Figure 3) helps validate the cluster separation.

[7]:

```r
# Now do the k-means clustering on all

data_for_clustering <- census_long %>%
  select(white, black, hisp, asian)

# Function to calculate total within-cluster sum of squares for different k
wss <- function(k) {
  kmeans(data_for_clustering, k, nstart = 10)$tot.withinss
}
```

```
# Compute and plot wss for k = 1 to k = 10
k.values <- 1:10
wss_values <- map_dbl(k.values, wss)

# Elbow method plot
plot(k.values, wss_values, type = "b", pch = 19, frame = FALSE,
     xlab = "Number of clusters K",
     ylab = "Total within-clusters sum of squares")
```

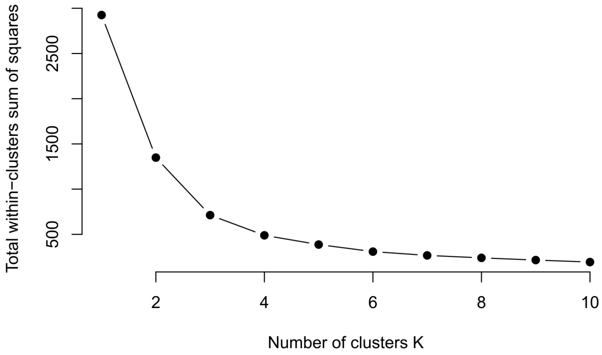

Figure 2: Elbow method plot – Clustering analysis

[7]:

```
Output in Figure 2
```

[8]:

```
# Silhouette method for determining the optimal number of clusters
fviz_nbclust(data_for_clustering, kmeans, method = "silhouette")
```

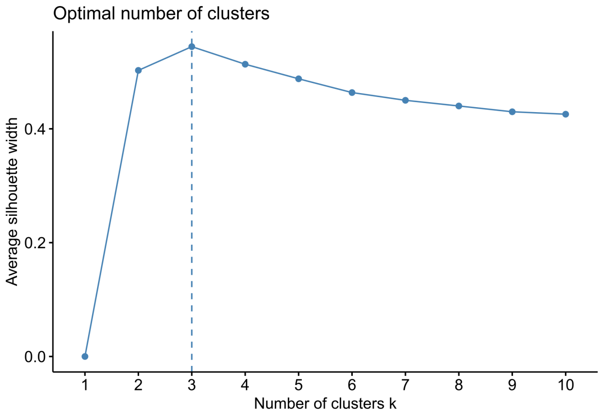

Figure 3: Silhouette method plot – Clustering analysis

[8]:

```
Output in Figure 3
```

5.2 Exploring k-means *Cluster Solutions*

According to these plots, the mathematically optimal number of neighborhood clusters for racial makeup is three. We can further explore the makeup of neighborhoods within these three clusters a few ways to determine if, in fact, three clusters provides a meaningful segmentation of four distinct racial and ethnic groups. In the plots below, we can visualize the average silhouette value for each cluster. The *Silhouette* values range from -1 to 1; values close to 1 suggest that the observations are well clustered while negative values suggest that an observation might be assigned to the wrong cluster. From the plot, Clusters 2 and 3 appears to be the most cohesive clusters, with average silhouette widths of 0.64, while cluster 1 has some potentially poorly classified neighborhoods. Descriptions of the racial and ethnic makeup of the clusters are obtained from the associated stacked

Table 1: Cluster Profiles: Average Demographics

Cluster	White	Black	Hispanic	Asian
1	0.15	0.18	0.50	0.15
2	0.07	0.75	0.14	0.02
3	0.73	0.04	0.12	0.10

bar charts. We can see that Cluster 1 is characterized as being nearly 50 percent Hispanic, with near equal shares of Whites, Blacks, and Asians. Cluster 2 is majority Black with approximately 15 percent Hispanics and few Whites and Asians. Finally, Cluster 3 is majority White. Therefore, this segmentation provides us with clusters indicating the dominant racial groups, but may miss some nuances of other racial and ethnic neighborhood compositions. We can therefore explore how increasing k may portray a richer portrait of neighborhood demographic profiles. Figure 4 illustrates the Silhouette Analysis to assess cluster separation, and the demographic makeup of the three clusters.

[9]:

```
# Assume the optimal number of clusters (k) is 3 from the previous steps
set.seed(123)
kmeans_result <- kmeans(data_for_clustering, centers = 3, nstart = 25)

# Add the cluster assignments to the original data
census_long$cluster <- kmeans_result$cluster

# Silhouette Analysis
sil <- silhouette(kmeans_result$cluster, dist(data_for_clustering))
fviz_silhouette(sil, print.summary=FALSE)
```

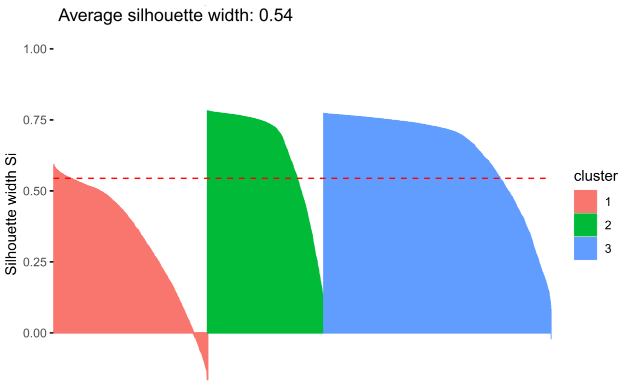

Figure 4: Cluster silhouette plot

[9]:

```
Output in Figure 4
```

[10]:

```
# Extract silhouette information to a data frame (df)
sil_df <- as.data.frame(sil[, 1:3])
colnames(sil_df) <- c("Cluster", "Silhouette Width", "Neighboring Cluster")

# Cluster profiles
cluster_profiles <- census_long %>%
  group_by(cluster) %>%
  summarise(across(c(white, black, hisp, asian), ~ round(mean(.), 2)))

# Print the table as a formatted table
kable(cluster_profiles, caption = "Cluster Profiles: Average Demographics",
      col.names = c("Cluster", "White", "Black", "Hispanic", "Asian"),
      format = "markdown")
```

[10]:

```
Output in Table 1
```

[11]:

```
# Reshape the data from wide to long format
cluster_profiles_long <- cluster_profiles %>%
  pivot_longer(cols = c("white", "black", "hisp", "asian"),
               names_to = "Demographic", values_to = "Proportion")

# Create the stacked bar chart
ggplot(cluster_profiles_long, aes(x = factor(cluster), y = Proportion,
                                  fill = Demographic)) +
  geom_bar(stat = "identity") +
  labs(title = "Cluster Demographic Makeup - 3 clusters", x = "Cluster",
       y = "Proportion") +
  scale_fill_brewer(palette = "Set3") +  # Adjust color palette if desired
  theme_minimal()
```

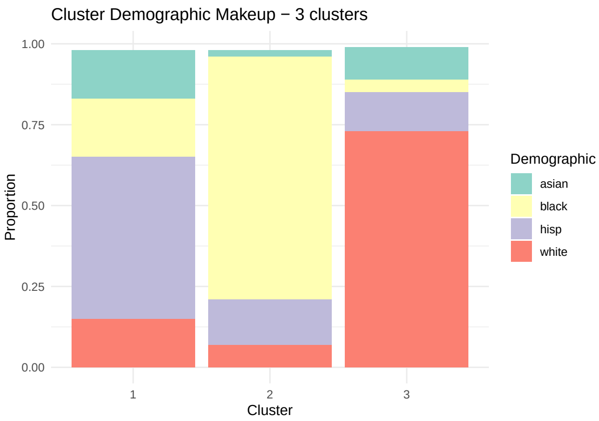

Figure 5: Cluster Analysis Results: Silhouette Analysis and Cluster Demographic Makeup

[11]:

```
Output in Figure 5
```

Next, we compare 4, 5, and 6 cluster solutions. We can see that the average silhouette of the solutions declines as the number of clusters increases. The demographic profiles show several new neighborhood typologies emerge with more clusters added. With a four cluster solution, we observe neighborhood typologies for each of the three dominant racial and ethic groups: White, Hispanic, and Black along with one mixed neighborhood type, shown in cluster 4. As expected, that cluster displays the lowest silhouette value, with some potentially mis-classified neighborhoods. The 5 cluster solution adds a cluster showing a majority Asian population, alongside two majority White populations - one showing more diversity than the other), and a majority Black, and Hispanic group. Finally, a 6 cluster solution shows more racially mixed groups, but at the expense of less well-defined or separated clusters. In the code below, each variable name ends with a number (4, 5 or 6), indicating which number of clusters (k) it represents. The results displays the silhouette analysis results for clustering into (a) four clusters, (b) five clusters, and (c) six clusters, illustrating the cohesion and separation of clusters at different sizes.

[12]:

```
# Perform k-means clustering for 4, 5, and 6 clusters
set.seed(123)
kmeans_4 <- kmeans(data_for_clustering, centers = 4, nstart = 25)
kmeans_5 <- kmeans(data_for_clustering, centers = 5, nstart = 25)
kmeans_6 <- kmeans(data_for_clustering, centers = 6, nstart = 25)

# Add cluster assignment to original data
census_long$cluster_4 <- kmeans_4$cluster
census_long$cluster_5 <- kmeans_5$cluster
census_long$cluster_6 <- kmeans_6$cluster

# Silhouette analysis for 4, 5, and 6 clusters
sil_4 <- silhouette(kmeans_4$cluster, dist(data_for_clustering))
sil_5 <- silhouette(kmeans_5$cluster, dist(data_for_clustering))
sil_6 <- silhouette(kmeans_6$cluster, dist(data_for_clustering))
```

```
# Plot silhouette for 4 clusters
plot_4 <- fviz_silhouette(sil_4, print.summary=FALSE) +
  ggtitle("Silhouette Plot - Four Clusters")

# Plot silhouette for 5 clusters
plot_5 <- fviz_silhouette(sil_5, print.summary=FALSE) +
  ggtitle("Silhouette Plot - Five Clusters")

# Plot silhouette for 6 clusters
plot_6 <- fviz_silhouette(sil_6, print.summary=FALSE) +
  ggtitle("Silhouette Plot - Six Clusters")

# Combine the three plots into a faceted view
grid.arrange(plot_4, plot_5, plot_6, ncol = 1)
```

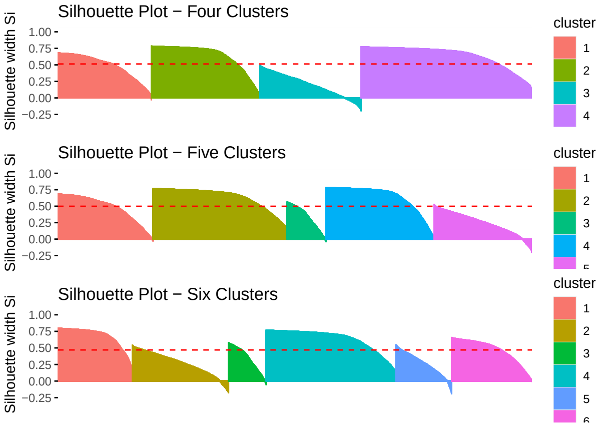

Figure 6: Silhouette Analysis for Different Cluster Sizes: (top) Four Clusters, (middle) Five Clusters, and (bottom) Six Clusters.

[12]:
```
Output in Figure 6
```

Figure 6 illustrates the demographic composition of clusters for different clustering solutions: (top) four clusters, (middle) five clusters, and (bottom) six clusters, highlighting how demographic groups are distributed across various cluster labels.

[13]:
```
# We need to hard code the labels to be able to compare the demographic profile
# of each solution. This is because R randomly assigns the label each time,
# even if the clustering solution is the same because we set the seed.

label_clusters <- function(cluster_profiles) {
  cluster_profiles %>%
    mutate(
      label = case_when(
        white > 0.75 ~ "3 White",
        hisp > 0.45 ~ "2 Hispanic",
        black > 0.75 ~ "1 Black",
        asian > 0.45 ~ "5 Asian",
        white < 0.56 & hisp > 0.20 & black < 0.20 ~ "4 Mixed",
        black > 0.49 & hisp > 0.25 & white < 0.15 ~ "6 Black and Hispanic",
        TRUE ~ "Other"  # Default label if none of the conditions are met
      )
    )
}

# Create cluster profiles and label them according to your custom rules
cluster_profiles_4 <- census_long %>%
  group_by(cluster_4) %>%
  summarise(across(c(white, black, hisp, asian), ~ round(mean(.), 2))) %>%
  mutate(`Clustering_Solution` = "4 Clusters", Cluster = cluster_4) %>%
  label_clusters()

cluster_profiles_5 <- census_long %>%
  group_by(cluster_5) %>%
  summarise(across(c(white, black, hisp, asian), ~ round(mean(.), 2))) %>%
  mutate(`Clustering_Solution` = "5 Clusters", Cluster = cluster_5) %>%
  label_clusters()

cluster_profiles_6 <- census_long %>%
  group_by(cluster_6) %>%
  summarise(across(c(white, black, hisp, asian), ~ round(mean(.), 2))) %>%
  mutate(`Clustering_Solution` = "6 Clusters", Cluster = cluster_6) %>%
  label_clusters()

# Combine all cluster profiles into one table
cluster_profiles_combined <-
  bind_rows(cluster_profiles_4, cluster_profiles_5, cluster_profiles_6)

# Map the labels back to the original tracts by joining based on cluster assignments
census_long <- census_long %>%
  left_join(cluster_profiles_4 %>% select(Cluster, label) %>% rename(label_4 = label),
            by = c("cluster_4" = "Cluster")) %>%
  left_join(cluster_profiles_5 %>% select(Cluster, label) %>% rename(label_5 = label),
            by = c("cluster_5" = "Cluster")) %>%
  left_join(cluster_profiles_6 %>% select(Cluster, label) %>% rename(label_6 = label),
            by = c("cluster_6" = "Cluster"))

# Reshape data for plotting
cluster_profiles_long <- cluster_profiles_combined %>%
  pivot_longer(cols = c("white", "black", "hisp", "asian"),
               names_to = "Demographic", values_to = "Proportion")

# Plot stacked bar charts of demographic makeup for each labeled cluster
ggplot(cluster_profiles_long, aes(x = label, y = Proportion, fill = Demographic)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_grid(Clustering_Solution ~ ., scales = "free_x") +
      # Facet vertically by clustering solution
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(
    title = "",
    x = "Cluster Label",
    y = "Proportion of Demographic Group",
    fill = "Demographic Group"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5)
  )
```

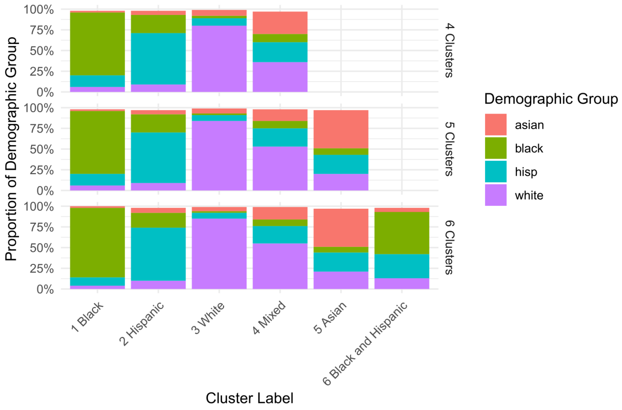

Figure 7: Geographic variation of Labeled Clusters for Different Clustering Solutions.

[13]:
```
Output in Figure 7
```

We can explore how this plays out for a specific observation. For example, take the first census tract in the data frame for the year 2020. This tract's racial composition was 58% Black and 30% Hispanic with small shares of Whites and Asians. With the four cluster solution, this tract was classified into class 1, `Majority Black`. For the five cluster solution, it was also classified as `Majority Black`, and for the six cluster solution, `Black and Hispanic`. In 1990, that same tract was 34% black and 61% Hispanic, resulting in the tract being classified as `Majority Hispanic` by all three clustering solutions. When looking at change over time, the neighborhood will either be registered as transitioning from `Majority Hispanic` to either `Majority Black` or `Black and Hispanic`, depending on the final cluster solution. In this instance, the 6 cluster solution provides a more accurate portrayal of the dynamics - while the neighborhood did technically become majority Black, there is still a significant Hispanic presence, a detail that would have been omitted by limiting the number of groups.

Finally, we can examine the spatial distribution of these clustering solutions to help aid in the final determination for the sequence analysis. Since we did the cluster analysis on all decades, for this purpose, we will contrast the 4, 5, and 6 cluster solutions just for 2020. All three maps depict a similar spatial pattern, but with a more fragmented pattern in the case of the 6 cluster solution. For example, the first two maps show contiguous tracts of the predominantly Black cluster, but the third map shows the emergence of the `Black and Hispanic` group largely forming on the outskirts of this spatial cluster. Another apparent distinction is the large spatial cluster of the `Asian` cluster in Queens, which had been labeled as `Mixed` in the 4 cluster solution.

To better understand details on racial neighborhood transitions, we will go with the larger number of clusters, 6, despite the mathematical preference for a 3 cluster solution. Our result illustrates the demographic distribution of labeled clusters for different clustering solutions: four clusters, five clusters, and six clusters.

[14]:

```r
# Filter data for the year 2020 and merge labels with tract data for mapping
tract_clusters_2020 <- tract %>% erase_water(area_threshold = 0.9) %>%
  left_join(census_long %>%
      filter(year == 2020) %>%  # Filter for the year 2020
      select(TRTID10, cluster_4, cluster_5, cluster_6, label_4, label_5, label_6),
            by = "TRTID10") %>%
  pivot_longer(cols = c("cluster_4", "cluster_5", "cluster_6"),
      names_to = "ClusterSolution", values_to = "Cluster") %>%
  pivot_longer(cols = c("label_4", "label_5", "label_6"),
      names_to = "LabelSolution", values_to = "Label") %>%
  filter(str_replace(ClusterSolution, "cluster_", "") == str_replace(LabelSolution,
                                                                  "label_", "")) %>%
  mutate(Label = factor(Label)) %>%
  mutate(ClusterSolution = recode(ClusterSolution,
                                  "cluster_4" = "4 Clusters",
                                  "cluster_5" = "5 Clusters",
                                  "cluster_6" = "6 Clusters")) %>%
  mutate(ClusterSolution = factor(ClusterSolution, levels = c("4 Clusters", "5 Clusters",
                                                           "6 Clusters")))

# Create the base plot with the labeled clusters
base_plot <- ggplot(tract_clusters_2020) +
  geom_sf(aes(fill = Label), color = NA) +  # Use the Label field for fill
  geom_sf(data = nyc_counties, fill = NA, color = "black", size = 0.5) +
    # Add county outlines
  facet_wrap(~ ClusterSolution, nrow = 1) +  # Facet horizontally by clustering solution
  scale_fill_manual(
    values = c("3 White" = "#F0E442", "2 Hispanic" = "#D55E00",
               "1 Black" = "#0072B2", "5 Asian" = "#CC79A7",
               "4 Mixed" = "#009E73", "6 Black and Hispanic" = "#56B4E9"),
    labels = c("Black", "Hispanic", "White", "Asian", "Mixed", "Black and Hispanic")
      # Remove numbers from labels
  ) +
  labs(x = "", y = "", fill = NULL) +
    # Remove the title and set fill to NULL to remove "Label"
  theme_void() +
  theme(
    strip.text = element_text(hjust = .5, vjust = .1, face = "italic", size = 12),
      # Center facet labels
    legend.position = "bottom",  # Place the legend at the bottom
    legend.title = element_blank(),  # Ensure legend title is blank
    legend.text = element_text(size = 8)
  ) +
  guides(fill = guide_legend(nrow = 1, byrow = TRUE, label.position = "bottom"))
    # Place category names under the boxes

# Display the plot
grid.arrange(base_plot)
```

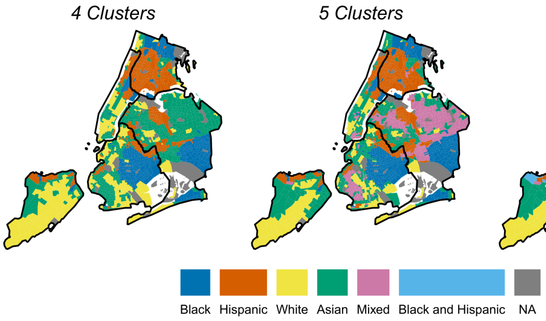

Figure 8: Demographic Distribution of Labeled Clusters for Different Clustering Solutions: Four Clusters (top), Five Clusters (middle), and Six Clusters (bottom).

[14]:

```
Output in Figure 8
```

We can begin with a simple exploration of the spatial changes over time in the maps showing the six clusters from 1980-2020 in a series of small multiples. From these maps, we can pick out some general spatial patterns over time. For example, we can see the expanse of neighborhoods classified as majority White from 1980 significantly diminishes by 2020. The share of Hispanics is shown to increase over time in the northern sections of the City. Towards the East, we see neighborhoods generally transition from predominantly White in 1980 to White and Mixed Race and eventually to Asian by 2020. There are also some evident stable clusters. For instance, the two clusters of predominantly black neighborhoods in the South and Southeast appear quite stable over time. However, the cluster of majority Black neighborhoods in towards the north of Manhattan is diminished, replaced by a Black and Hispanic classification.

[15]:

```
decades <- c(1980, 1990, 2000, 2010, 2020)

# Filter and prepare data for mapping
tract_clusters_decades <- tract %>%
  left_join(census_long %>%
              select(TRTID10, year, label_6), by = "TRTID10") %>%
  filter(!is.na(label_6))

# Set factor levels for year to ensure proper ordering
tract_clusters_decades <- tract_clusters_decades %>%
  mutate(year = factor(year, levels = decades))

# Plot faceted maps for each decade (year)
ggplot(tract_clusters_decades) +
  geom_sf(aes(fill = label_6), color = NA) +
    # Use factor to ensure proper color assignment
  geom_sf(data = nyc_counties, fill = NA, color = "black", size = 0.5) +
    # Add county outlines
  facet_wrap(~ year, ncol = 5) +  # Facet by year with 5 columns
  scale_fill_manual(
    values = c("3 White" = "#F0E442", "2 Hispanic" = "#D55E00", "1 Black" = "#0072B2",
               "5 Asian" = "#CC79A7", "4 Mixed" = "#009E73",
               "6 Black and Hispanic" = "#56B4E9"),
    labels = c("Black", "Hispanic", "White", "Asian", "Mixed", "Black and Hispanic"),
      # Remove numbers from labels
    name = ""  # Remove legend title
  ) +
  labs(title = " ", x = "", y = "") +
  theme_void() +
  theme(
    plot.title = element_text(hjust = 0, color = "gray"),  # Set the title color to gray
    legend.position = "bottom",  # Move legend to the bottom
    legend.direction = "horizontal",  # Make the legend horizontal
    legend.title = element_blank(),  # Ensure the legend title is blank
    legend.key.size = unit(0.6, "cm"),  # Set consistent size for legend boxes
    legend.key.height = unit(0.6, "cm"),  # Set consistent height for legend boxes
    legend.key.width = unit(0.5, "cm")  # Set consistent width for legend boxes
  ) +
  guides(
    fill = guide_legend(
      nrow = 1, byrow = TRUE, label.position = "bottom"
        # Place category names under the legend boxes
    )
  )
```

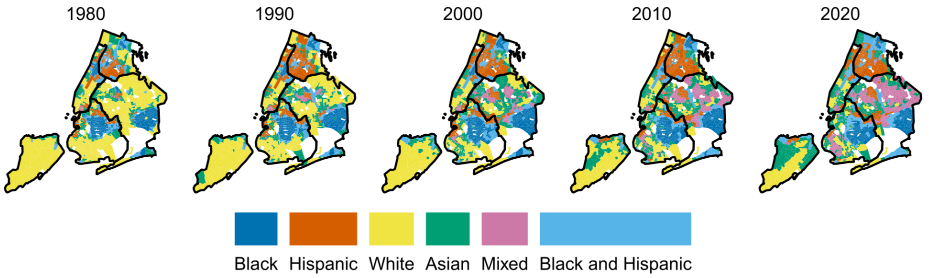

Figure 9: Change in the cluster membership over five decades

[15]:
```
Output in Figure 9
```

5.3 *Sequence Analysis*

Our objective with the sequence classification is to come up with a typology of neighborhood sequences over time to describe general pathways of change. Each neighborhood has a sequence of classes over time, one of 6 categorical groups for each of the five decennial census values from 1980-2020. The first step in the analysis is to convert the data frame into a sequence for each neighborhood. We can see an illustrative example of the longitudinal sequences from the first five records. Classes are separated by a dash (-). There are 2122 sequences in the dataset and 197 unique sequences; all sequences are displayed in the plot below. Thus, the purpose of clustering the sequences is to extract the general patterns present from the set of all sequences.

[16]:
```
# Convert the data from long to wide format to have a sequence for each neighborhood
# Assuming census_long is your data frame
census_wide <- tract_clusters_decades %>%
  st_drop_geometry() %>%
  select(TRTID10, year, label_6) %>%
  pivot_wider(names_from = year, values_from = label_6)

# Rename the columns to show only the year (remove "cluster_" prefix if it was added
# during previous steps)
colnames(census_wide) <- sub("cluster_", "", colnames(census_wide))

# Ensure columns are in the correct order by year
census_wide <- census_wide %>%
  select(TRTID10, `1980`, `1990`, `2000`, `2010`, `2020`)

# Ensure the sequence columns are factors
census_wide <- census_wide %>%
  mutate(across(starts_with("19") | starts_with("20"), as.factor))

# Check the distinct states (categories) in your sequences
unique_states <- unique(unlist(census_wide[, -1]))  # Exclude TRTID10 column
num_states <- length(unique_states)
print(unique_states)
```

[16]:

```
[1] 6 Black and Hispanic 2 Hispanic           4 Mixed
[4] 3 White              1 Black              5 Asian
Levels: 1 Black 2 Hispanic 3 White 4 Mixed 5 Asian 6 Black and Hispanic
```

[17]:

```
print(num_states)  # Number of unique states
```

[17]:

```
[1] 6
```

[18]:

```
# Define the custom color palette with the correct colors
custom_palette <- c(
  "1 Black" = "#0072B2",
  "2 Hispanic" = "#D55E00",
  "3 White" = "#F0E442",
  "4 Mixed" = "#009E73",
  "5 Asian" = "#CC79A7",
  "6 Black and Hispanic" = "#56B4E9"
)

# Create the sequence object with the renamed columns
sequence_data <- seqdef(census_wide[, -1], cpal = custom_palette)
# Exclude the TRTID10 column

# Check the number of distinct sequences
num_sequences <- seqtab(sequence_data, idx = 0) %>% nrow
print(num_sequences)
```

[18]:

```
[1] 197
```

[19]:

```
# Display the first few sequences
head(sequence_data)
```

[19]:

```
  Sequence
1 6 Black and Hispanic-2 Hispanic-6 Black and Hispanic-6 Black and Hispanic-6 Black and
                                                                              Hispanic
2 2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic
3 2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic
4 2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic
5 2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic-2 Hispanic
6 4 Mixed-2 Hispanic-2 Hispanic-2 Hispanic-5 Asian
```

[20]:

```
# Plot the sequences
seqIplot(sequence_data,                         # Sequence object
         with.legend = "right",                 # Display legend on right side of plot
         cex.legend = 0.6,                      # Change size of legend
         main = "Neighborhood Racial and Ethnic Trajectories") # Plot title
```

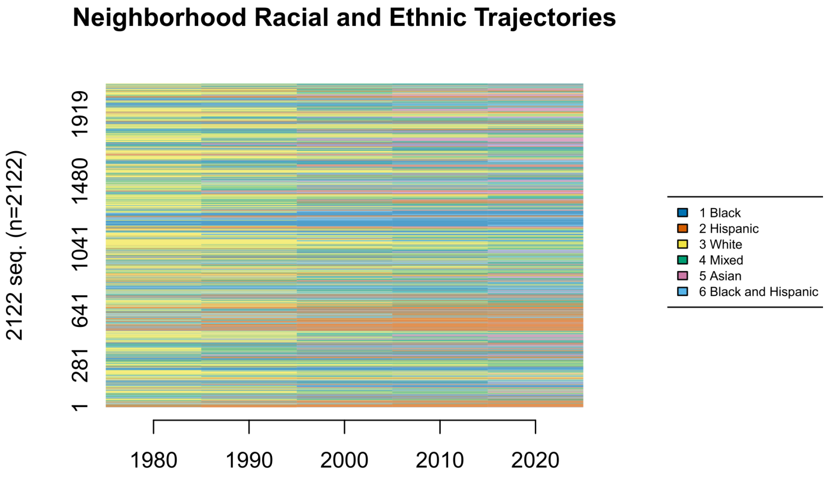

Figure 10: Sequence of each neighborhood.

[20]:

```
Output in Figure 10
```

The next objective is to compute the dissimilarity between all sequences. As described previously, we use the OMstrans algorithm to preserve the ordering of events as we are most interested in describing how neighborhoods have generally transitioned over time. We set a low value (`0.1`) for the parameter `otto` in the `seqdist` command which is the origin-transition trade-off weight. This emphasizes the ordering of sequence states. We also set a high `indel` cost of `3`. Finally, substitution costs are a function of the transition rate `TRATE` between states. This places a lower cost on more frequent transitions.

The substitution cost matrix is shown in Table 2. The table indicates that more frequent transitions from, for example, Black & Hispanic to Black, Mixed to White, White to Mixed, and Black to Black and Hispanic have lower costs than rarer transitions from, for example, Asian to Black, Hispanic to Black, or Black to Asian. Generally, the more mixed race groups are more transitional white the majority race groups tend to transition through a mixed race state. This will become more evident as we examine the resulting sequence clusters.

Once the cost matrix is established, we then cluster the sequences using the dissimilarity matrix as an input to generate the typology. We follow an iterative process in determining the optimal solution like the one described above for the k means clustering. In short, multiple solutions are tested and the resulting sequence clusters are visualized to inspect for heterogeneity. Because our clustering and distance matrix are optimized for describing transitions over time, we end up with one cluster that contains all neighborhoods that remained constant over time. These sequences are very different from one another in the neighborhood types they describe, but they all represent a pathway or sequences of no change.

To describe the resulting sequence clusters, we plot a Sequence Frequency Plot for each cluster. We settled on 14 trajectory clusters describing neighborhood racial and ethnic transitions from 1980 to 2020 in New York City. The Sequence Frequency Plots illustrate the sequences belonging to each cluster and the sequence bars are scaled to visualize the frequency of each sequence. A summary of the trajectories is as follows:

1. Hispanic Majority to Black and Hispanic
2. Stability Cluster
3. White and Mixed Race to Hispanic Majority
4. White to White and Mixed Race to Hispanic Majority
5. Black and Hispanic to Hispanic Majority
6. White Majority to White and Mixed Race
7. White and Mixed Race to Black and Hispanic
8. Black and Hispanic to Black Majority
9. White and Mixed Race to Asian Majority
10. White and Mixed Race to White Majority
11. Black and Hispanic to White and Mixed Race
12. White and Mixed Race to Asian Majority
13. Black Majority to Black and Hispanic
14. Black and Hispanic to Asian Majority

[21]:

```r
# Pass this palette to seqdef
# Define sequence data with custom color palette
sequence_data <- seqdef(census_wide[, -1], cpal = custom_palette)
  # Exclude the TRTID10 column

# Compute Optimal Matching (OM) distances using the TRATE cost method
costs <- seqcost(sequence_data, method = "TRATE")
om_distances <- seqdist(sequence_data, method = "OMstran", indel = 3, sm = costs$sm,
                        otto = 0.1)

#Extract the substitution cost matrix
sub_matrix <- round(costs$sm, 2)  # round to 2 decimal places for clarity

# Convert to a data frame for export
sub_df <- as.data.frame(sub_matrix)

# Add row names as a column for better display
sub_df$From <- rownames(sub_df)
sub_df <- sub_df[, c("From", setdiff(names(sub_df), "From"))]

# Optional: reorder columns to match row order
sub_df <- sub_df[, c("From", rownames(sub_matrix))]
```

Table 2: Substitution Cost Matrix

From	1 Black	2 Hispanic	3 White	4 Mixed	5 Asian	6 Black & Hisp
1 Black	0.00	2.00	1.99	1.99	2.00	1.77
2 Hispanic	2.00	0.00	2.00	1.86	1.94	1.86
3 White	1.99	2.00	0.00	1.70	2.00	1.99
4 Mixed	1.99	1.86	1.70	0.00	1.80	1.91
5 Asian	2.00	1.94	2.00	1.80	0.00	1.97
Black & Hisp	1.77	1.86	1.99	1.91	1.97	0.00

```
# Output to table
sub_df[6,1] <- "Black & Hisp"
names(sub_df)[names(sub_df) == '6 Black and Hispanic'] <- '6 Black & Hisp'
knitr::kable(sub_df, row.names=FALSE)
```

[21]:

```
Output in Table 2
```

[22]:

```
# Perform hierarchical clustering using Ward's method on the OM distances
clusterward <- agnes(om_distances, diss = TRUE, method = "ward")

# Define the number of clusters and assign clusters to the sequence data
num_clusters <- 14
clusters <- cutree(clusterward, k = num_clusters)
census_wide$sequence_cluster <- clusters

# Define cluster names for reference
cluster_names1 <- c(
  "1 Black & Hispanic to Hispanic Majority",
  "2 Stability",
  "3 White Mixed Race to Hispanic",
  "4 White to Mixed Race to Hispanic Majority",
  "5 Majority White to Increasing Diversity"
)

# Plot sequences for clusters in groups, adjusting legend settings
plot_sequence_clusters <- function(cluster_range) {
  seqfplot(
   sequence_data[census_wide$sequence_cluster %in% cluster_range, ],
   group = census_wide$sequence_cluster[census_wide$sequence_cluster %in% cluster_range],
   sortv = "from.start",
   border = NA,
   with.legend = "right",         # Place the legend on the right
   legend.prop = 0.2,             # Set the proportion of the plot area for the legend
   legend.border = FALSE          # Remove the border around the legend
  )
}

# Plot sequences for specified cluster ranges
plot_sequence_clusters(1:2)
```

[22]:

```
Output in Figure 11
```

[23]:

```
plot_sequence_clusters(3:4)
```

[23]:

```
Output in Figure 12
```

[24]:

```
plot_sequence_clusters(5:6)
```

[24]:

```
Output in Figure 13
```

[25]:

```
plot_sequence_clusters(7:8)
```

[25]:

```
Output in Figure 14
```

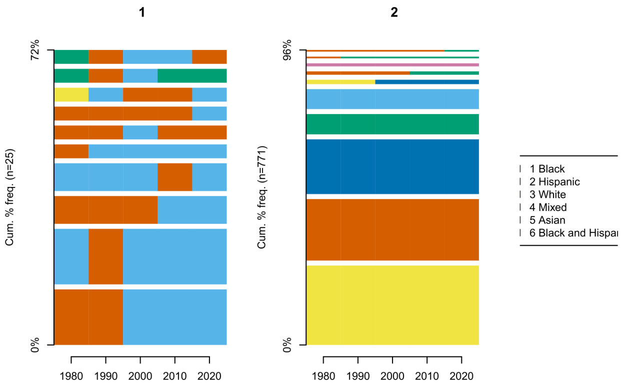

Figure 11: Sequences for Clusters 1 and 2

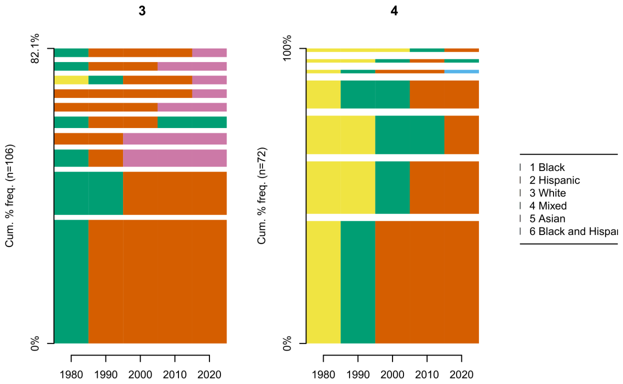

Figure 12: Sequences for Clusters 3 and 4

```
[26]: plot_sequence_clusters(9:10)
```

```
[26]: Output in Figure 15
```

```
[27]: plot_sequence_clusters(11:12)
```

```
[27]: Output in Figure 16
```

```
[28]: plot_sequence_clusters(13:14)
```

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

Of these 14 pathways, there are 3 that lead to the formation of a neighborhood

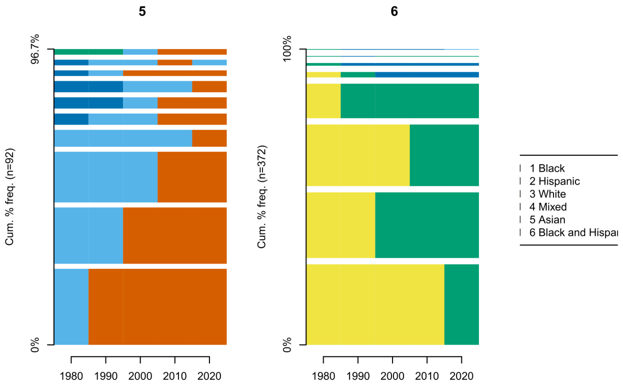

Figure 13: Sequences for Clusters 5 and 6

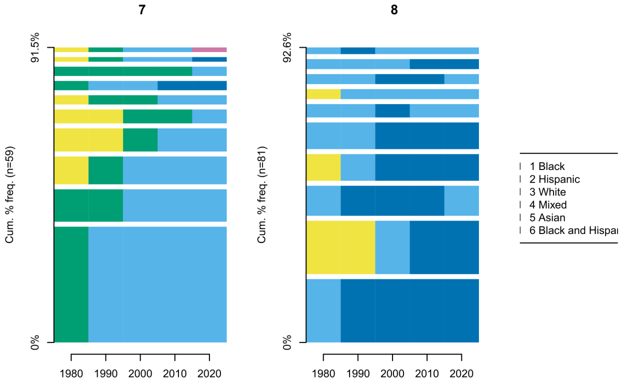

Figure 14: Sequences for Clusters 7 and 8

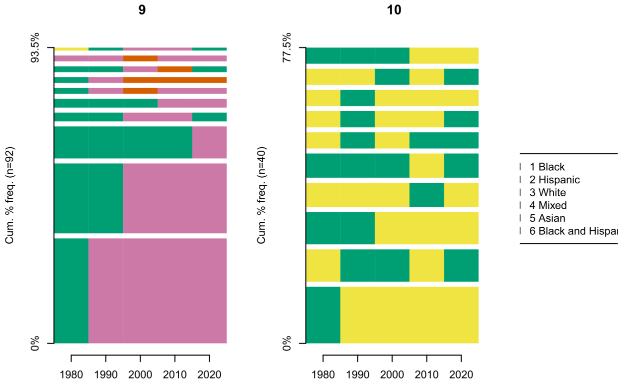

Figure 15: Sequences for Clusters 9 and 10

transitioning into a Hispanic Majority cluster by 2020. This includes Cluster 3 - showing neighborhoods that went from being a slight majority White, but with a mixture of other races in 1980 and 1990 to transitioning to majority Hispanic by 1990 or 2000. Some of these sequences indicate a continued transition towards becoming majority Asian in the later years. Further segmenting the sequences into more clusters may have separated out those trajectories, but for the sake of brevity, we leave them mixed in. Spatially, these are shown in orange on the map below and can be see clustered in Staten Island, Queens, and in the Northern portion of the city. They are also notably adjacent to neighborhoods indicated by the red color, those representing sequence cluster 4, transitioning from majority White to White and Mixed Race and then to Majority Hispanic. This latter cluster might represent the precursor to cluster 3, but are neighborhoods that made this transition from majority White later, where the changes took time to spatially spillover

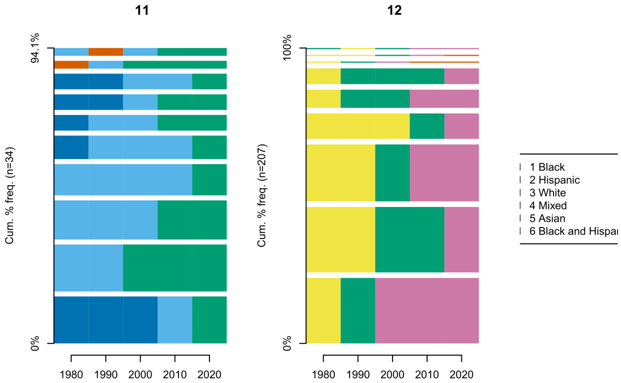

Figure 16: Sequences for Clusters 11 and 12

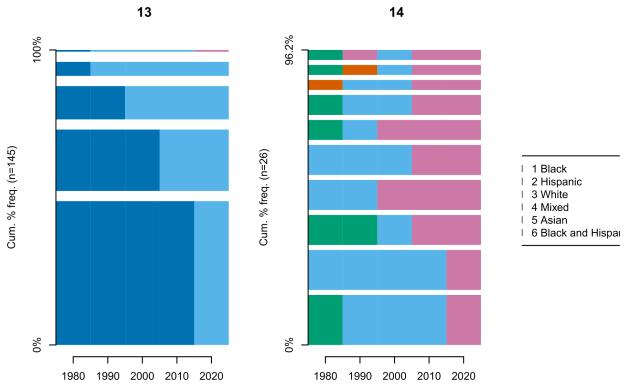

Figure 17: Sequences for Clusters 13 and 14

to adjacent neighborhoods, as indicated by the map. Both of these pathways depict a transition from largely White populations to largely Hispanic.

The third pathway depicting a transition to majority Hispanic is distinct. It is represented by cluster 5 showing a transition from either majority Black neighborhoods towards a mixed Hispanic and Black group and eventually majority Hispanic or, beginning the 1980 time stamp, in a more mixed Black and Hispanic state. Geographically, these neighborhoods are shown more in the northern sections of Manhattan and the Bronx.

From these two sets of sequence clusters, we can see that majority Hispanic neighborhoods in New York City have emerged out of either majority Black or Majority White neighborhoods over time.

[29]:
```r
# Define the Hispanic majority clusters for plotting
hispanic_majority_clusters <- census_wide %>%
  filter(sequence_cluster %in% c(4, 5, 6))  # Filter clusters 4, 5, and 6

# Join the filtered data with the tract shapefile based on TRTID10
tract_hispanic_majority <- tract %>%
  left_join(hispanic_majority_clusters, by = "TRTID10") %>%
  erase_water(area_threshold = 0.75)

# Add county borders and color to the plot
ggplot(tract_hispanic_majority) +
  geom_sf(aes(fill = factor(sequence_cluster)), color = NA) +  # Map sequence clusters
  geom_sf(data = nyc_counties, fill = NA, color = "black", size = 0.5) +
    # Add white county borders
  scale_fill_manual(values = c("#FB8072", "#80B1D3", "#FDB462"),  # Custom color palette
                    labels = c("3 White - White, Mixed Race - Hispanic",
                               "5 Black - Black & Hispanic - Hispanic",
                               "4 White, Mixed Race - Hispanic"),
                    name = "Hispanic Majority Pathway") +
  labs(title = "",
       x = "", y = "") +
  theme_void() +
  theme(legend.position = "right",  # Place legend on the right
        legend.direction = "vertical",  # Arrange legend items vertically
        legend.title = element_text(size = 12),  # Customize legend title size
        legend.key.width = unit(2, "cm"),  # Adjust legend key width
        legend.box = "vertical")  # Place legend title on top of the legend
```

There are also 3 pathways leading to an Asian majority neighborhood type. These include clusters 9 and 12 which are also likely continuations of longer trajectories, that show a gradual transition from Majority White to White mixed race and eventually to Asian Majority. Geographically, these are clustered in the northern section of Queens. Sequence cluster 14 is more distinct in that the Asian majority transitioned from the Black and Hispanic mixed group and spatially, they are generally located in upper Manhattan and the Bronx.

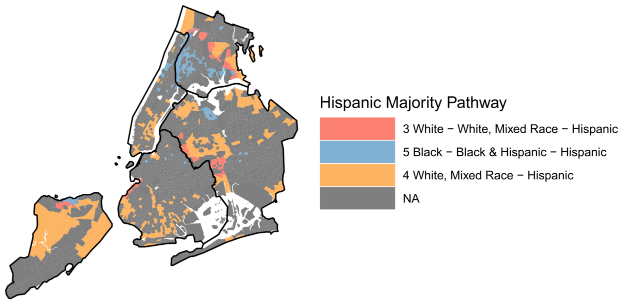
Hispanic Majority Pathway
3 White − White, Mixed Race − Hispanic
5 Black − Black & Hispanic − Hispanic
4 White, Mixed Race − Hispanic
NA

Figure 18: Neighborhoods Following Pathway to Hispanic Majority with White County Borders.

[30]:

```r
# Filter census data to include only clusters 14, 12, and 9
# (Asian Majority Pathway clusters)

asian_majority_clusters <- census_wide %>%
  filter(sequence_cluster %in% c(14, 12, 9))

# Join the filtered data with the tract shapefile (here, use TRTID10)
tract_asian_majority <- tract %>% erase_water(area_threshold = 0.75) %>%
  left_join(asian_majority_clusters, by = "TRTID10")  # Join with spatial data

# Create a map of the neighborhoods with clusters 14, 12, and 9 with new labels
ggplot(tract_asian_majority) +
  geom_sf(aes(fill = factor(sequence_cluster)), color = NA) +  # Map sequence clusters
  geom_sf(data = nyc_counties, fill = NA, color = "black", size = 0.5) +
    # Add white county borders
  scale_fill_manual(values = c("#8DD3C7", "#FFFFB3", "#BEBADA"),
    # Custom color palette for Asian Majority clusters
            labels = c("14 White, Mixed Race to Black & Hispanic to Asian Majority",
                       "12 White Majority to White Mixed Race to Asian",
                       "9 White Mixed Race to Asian"),
            name = "Asian Majority Pathway") +
  labs(title = "",
       x = "", y = "") +
  theme_void() +
  theme(legend.position = "right",  # Place legend at the bottom
        legend.direction = "vertical",  # Arrange legend items horizontally
        legend.title = element_text(size = 12),  # Customize legend title size
        legend.key.width = unit(2, "cm"),  # Adjust legend key width
        legend.box = "vertical")  # Place legend title on top of the legend
```

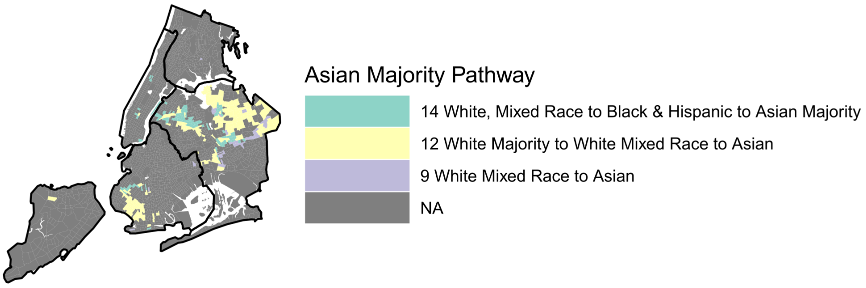
Asian Majority Pathway
14 White, Mixed Race to Black & Hispanic to Asian Majority
12 White Majority to White Mixed Race to Asian
9 White Mixed Race to Asian
NA

Figure 19: Neighborhoods Following Pathway to Asian Majority

[30]:

```
Output in Figure 19
```

[31]:

```r
# Filter data for Increasing White (clusters 10 and 11)
increasing_white_clusters <- census_wide %>%
  filter(sequence_cluster %in% c(10, 11))  # Clusters 10 and 11 for increasing White

# Filter data for Increasing Black (clusters 1 and 8)
increasing_black_clusters <- census_wide %>%
  filter(sequence_cluster %in% c(1, 8))  # Clusters 1 and 8 for increasing Black

# Join the filtered data with the tract shapefile for each group
tract_increasing_white <- tract %>%  erase_water(area_threshold = 0.75) %>%
  left_join(increasing_white_clusters, by = "TRTID10")
    # Join spatial data for increasing White

tract_increasing_black <- tract %>%   erase_water(area_threshold = 0.75) %>%
  left_join(increasing_black_clusters, by = "TRTID10")
    # Join spatial data for increasing Black
```

There are two pathways for increasing both Black and White shares in a neighborhoods. Neighborhoods that became increasingly White either followed a trajectory from White mixed race to majority White (Cluster 10) or from either all Black or Black and Hispanic to White and Mixed race (11). Notably, this transition largely took place within the past 1-2 decades, aligning with when gentrification trends became accentuated in some cities, including New York. We see a clear cluster of this latter group in Brooklyn, a borough whose gentrification trends have been well documented (Chronopoulos 2020, Halasz 2023).

For the case of increasing Black populations, Cluster 1 shows a pathway from Hispanic majority to mixed Black and Hispanic and Cluster 8 shows a gradual transition from Black and Hispanic to majority Black; a trend that largely begin towards the middle of the study period, around the 2000 census data mark. Spatially, the former is more dispersed, while the latter is depicted more clearly in the southern neighborhoods of Brooklyn.

[32]:

```
# Adjust plotting settings to improve map visibility
# Create the map for Increasing White, zoomed in on the five boroughs
ggplot(tract_increasing_white) +
  geom_sf(aes(fill = factor(sequence_cluster)), color = NA) +  # Map sequence clusters
  geom_sf(data = nyc_counties, fill = NA, color = "black", size = 0.5) +
    # Add white county borders
  scale_fill_manual(values = c("#8DD3C7", "#FFFFB3"),           # Custom color palette
            labels = c("White Mixed Race to Majority White",    # for White clusters
                       "Black and Hispanic to White Mixed Race"),
            name = "Increasing White Pathway") +
  coord_sf(xlim = c(-74.3, -73.7), ylim = c(40.5, 40.9), expand = FALSE) +
    # Zoom into the NYC area
  theme_void() +
  theme(legend.position = "right",  # Place legend at the right
        legend.direction = "vertical",  # Arrange legend items vertically
        legend.title = element_text(size = 12),  # Customize legend title size
        legend.key.width = unit(2, "cm"),
        legend.box = "vertical")  # Place legend title on top
```

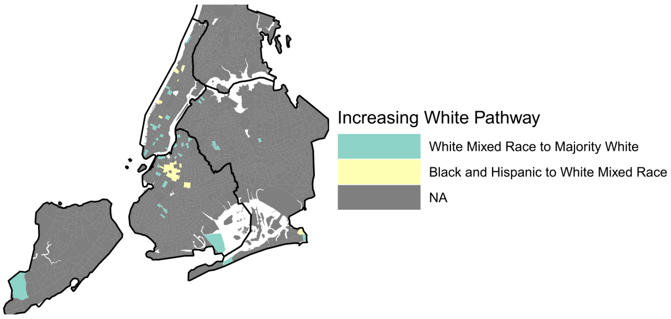
Increasing White Pathway
White Mixed Race to Majority White
Black and Hispanic to White Mixed Race
NA

Figure 20: Neighborhoods Following Pathway to White Majority

[32]:

```
Output in Figure 20
```

[33]:

```
# Create the map for Increasing Black, zoomed in on the five boroughs
ggplot(tract_increasing_black) +
  geom_sf(aes(fill = factor(sequence_cluster)), color = NA) +  # Map sequence clusters
  geom_sf(data = nyc_counties, fill = NA, color = "black", size = 0.5) +
    # Add white county borders
  scale_fill_manual(values = c("#FB8072", "#80B1D3"),           # Custom color palette
            labels = c("Hispanic Majority to Black & Hispanic", # for Black clusters
                       "Black and Hispanic to Black Majority"),
            name = "Increasing Black Pathway") +
  coord_sf(xlim = c(-74.3, -73.7), ylim = c(40.5, 40.9), expand = FALSE) +
    # Zoom into the NYC area
  theme_void() +
  theme(legend.position = "right",  # Place legend at the right
        legend.direction = "vertical",  # Arrange legend items vertically
        legend.title = element_text(size = 12),  # Customize legend title size
        legend.key.width = unit(2, "cm"),
        legend.box = "vertical")  # Place legend title on top
```

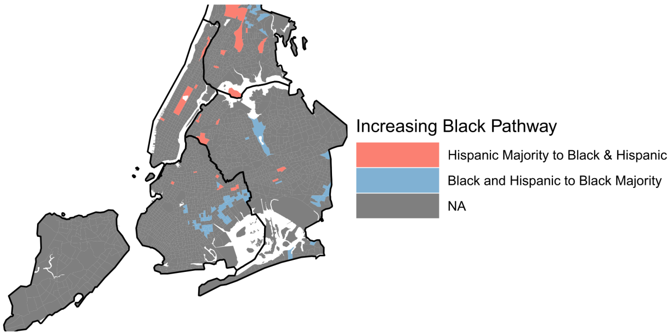
Increasing Black Pathway
Hispanic Majority to Black & Hispanic
Black and Hispanic to Black Majority
NA

Figure 21: Neighborhoods Following Pathway to Black Majority

[33]:
```
Output in Figure 21
```

Finally, of the sequences depicting no change from 1980 to 2020, shown in Cluster 2, the majority of those are for tracts with a racial majority: White, Hispanic, and Black. But those are followed by two stable sequences of neighborhoods with some racial diversity. The first is the majority White, but with a mixture of other races and the second is the mixed Black and Hispanic cluster. There is some debate in the literature whether racially mixed neighborhoods can be stable over time, or are they simply depicting a point along a pathway of change when one group will eventually become a majority. Research has shown that highly diverse neighborhoods are quite unstable over time, likely to transition to a less diverse state. In particular the transition from predominantly White towards majority Hispanic, results in a period of unstable racial mixture while that transition takes place (Wright et al. 2020). Others have identified a small, but persistent set of racially diverse neighborhoods throughout the United States, but particularly in racially diverse metropolitan areas (Hipp, Kim 2023). Here, we find some evidence of stable, non-racially homogeneous Census Tracts.

6 Conclusions

This analysis showcases a method for developing and visualizing a typology of neighborhood change pathways. We used a case study of decennial racial and ethnic changes in New York City census tracts from 1980-2020. The workflow first involves developing a cross-sectional typology of classes describing the racial mixture of neighborhoods. To do so, we used an unsupervised classification approach, k-means to derive six such clusters.

We demonstrated how determining the number of clusters often falls to more of an art than a precise science as our final clustering for this analysis exceeded the mathematically optimal three clusters, but provided us with more nuance on the racial mixture of neighborhoods. We performed the cluster analysis on all census tracts in the city for each of the five decennial census time stamps at once to ensure a temporally consistent set of groupings. We then created sequences of neighborhood clusters over the study period and developed a typology of sequences that grouped them based on the similarity of how they changed over time. To do so, we used the OMstrans algorithm for computing sequence dissimilarity to ensure that the ordering or sequencing of events was preserved. Finally, we mapped sequence clusters to spatially visualize neighborhoods that followed similar pathways of change. For our case study of the largest city in the United States, and one of the most diverse, we observed a decline in the number of majority White census tracts over time. We identified several pathways of change leading to majority Hispanic neighborhoods - emerging either out of previously majority White or Black neighborhoods. We also observed pathways leading to majority Asian tracts, transitioning from Black and Hispanic neighborhoods, largely in Queens. A smaller share of neighborhoods became either increasingly White or Black. Neighborhoods that saw an increase in the share of Whites, saw a notable increase in the transition from Black and Hispanic to White and mixed race over the past two decades. Neighborhoods that increased in the share of Blacks largely transitioned from Black and Hispanic to majority Black or from White and mixed race to Black and Hispanic.

The strength of the sequence analysis technique lies in its ability to clearly visualize common pathways of neighborhood change. One of its limitations, however, is the need to segment continuous, longitudinal data into discrete, categorical states. In this article, we devoted attention to this step, illustrating the use of a k-means algorithm as an unsupervised classification method for grouping multiple variables. This phase involves decisions, particularly, the number of classes, as that directly influences the resulting sequences and subsequent interpretations. Alternative methods for clustering time series data that preserve the continuous nature of the data may be preferable in cases where the number of neighborhood variables is limited (Delmelle et al. 2025).

Beyond describing trajectories, sequence analysis also opens up avenues for further inquiry. The resulting sequence clusters can serve as the basis for additional analyses like modeling the predictors of specific trajectories or analyzing their spatial patterns using categorical spatial autocorrelation measures such as the join-count statistic.

References

Chapple K, Zuk M (2016) Forewarned: The use of neighborhood early warning systems for gentrification and displacement. *Cityscape* 18: 109–130

Chronopoulos T (2020) What's happened to the people? Gentrification and racial segregation in Brooklyn. *Journal of African American Studies* 24: 549–572. CrossRef

Delmelle E (2015) Five decades of neighborhood classifications and their transitions: A comparison of four US cities, 1970–2010. *Applied Geography* 57: 1–11. CrossRef

Delmelle E (2017) Differentiating pathways of neighborhood change in 50 US metropolitan areas. *Environment and planning A* 49: 2402–2424. 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: 36–56. CrossRef

Delmelle EC (2022) GIScience and neighborhood change: Toward an understanding of processes of change. *Transactions in GIS* 26: 567–584. CrossRef

Delmelle EC, Nilsson I, Duma N (2025) Time series clustering for exploring neighborhood dynamics: The case of US neighborhood racial and ethnic trends, 1990–2020. *Geographical Analysis*. CrossRef

Farrell CR, Lee BA (2011) Racial diversity and change in metropolitan neighborhoods. *Social Science Research* 40: 1108–1123. CrossRef

Frey WH (2022) A new great migration is bringing black Americans back to the South. Brookings institution, https://www.brookings.edu/research/a-new-great-migration-is-bringing-black-americans-back-to-the-south/

Gabadinho A, Ritschard G, Müller NS, Studer M (2011) Analyzing and visualizing state sequences in R with TraMineR. *Journal of statistical software* 40: 1–37. CrossRef

Galster G (2001) On the nature of neighbourhood. *Urban Studies* 38: 2111–2124. CrossRef

González-Leonardo M, Newsham N, Rowe F (2023) Understanding population decline trajectories in Spain using sequence analysis. *Geographical Analysis* 55: 495–516. CrossRef

Halasz JR (2023) Between gentrification and supergentrification: Hybrid processes of socio-spatial upscaling. *Journal of Urban Affairs* 45: 771–796. CrossRef

Hipp JR, Kim JH (2023) Persistent racial diversity in neighborhoods: What explains it and what are the long-term consequences? *Urban Geography* 44: 640–667. CrossRef

Landis JD (2016) Tracking and explaining neighborhood socioeconomic change in US metropolitan areas between 1990 and 2010. *Housing Policy Debate* 26: 2–52. CrossRef

Logan JR, Xu Z, Stults BJ (2014) Interpolating US decennial census tract data from as early as 1970 to 2010: A longitudinal tract database. *The Professional Geographer* 66: 412–420. CrossRef

Patias N, Rowe F, Cavazzi S (2020) A scalable analytical framework for spatio-temporal analysis of neighborhood change: A sequence analysis approach. In: *Geospatial Technologies for Local and Regional Development: Proceedings of the 22nd AGILE Conference on Geographic Information Science 22*, 223–241. Springer

Reibel M, Regelson M (2011) Neighborhood racial and ethnic change: The time dimension in segregation. *Urban Geography* 32: 360–382. 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: 481–511. CrossRef

Terbeck FJ (2023) The impact of regional and local population trends on suburban poverty and ethnoracial composition change: A shift-share analysis of the Chicago metropolitan area in the 2000s. *Population, Space and Place* 28: e2549. CrossRef

Von Luxburg U, Williamson RC, Guyon I (2012) Clustering: Science or art? In: *Proceedings of ICML workshop on unsupervised and transfer learning*, 65–79. JMLR Workshop and Conference Proceedings

Wright R, Ellis M, Holloway SR, Catney G (2020) The instability of highly racially diverse residential neighborhoods in the United States. *Sociology of Race and Ethnicity* 6: 365–381. CrossRef

Wright R, Ellis M, Holloway SR, Wong S (2014) Patterns of racial diversity and segregation in the United States: 1990–2010. *The Professional Geographer* 66: 173–182. CrossRef

Zwiers M, van Ham M, Manley D (2018) Trajectories of ethnic neighbourhood change: Spatial patterns of increasing ethnic diversity. *Population, Space and Place* 24: e2094. CrossRef

Funded by

ersa

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

FWF