Assessing the spatial scale of segregation in the Netherlands.

Lucas Spierenburg *Corresponding author*
Dr. Oded Cats
Dr. Sander van Cranenburgh

1 Introduction

Spatial segregation, here understood as the uneven distribution of social groups in space (Reardon and O'Sullivan, 2004), is a persisting problem in many cities in the world (Tammaru et al., 2015; Wang et al., 2018). It can occur along one or several social dimensions, such as income, religion, or migration background. This situation is prejudicial for society, as segregation can result in exacerbating inequality between groups; in terms of education achievements, well-being, or health condition, among other aspects of people's life (Owens, 2018; Ludwig et al., 2012; Williams and Collins, 2001).

The spatial scale at which such segregation unfolds matters: in a city with large segregated regions, individuals from different groups are distant from each other, and thus less likely to encounter. This impedes interaction between groups, which is found to further contribute to inequality (Wilson, 1987; Vervoort, 2012; Tóth et al., 2021). Several studies have proposed methods to determine the spatial scale of segregation, and the factors influencing it (Reardon et al., 2008; Petrović et al., 2018; Veneri et al., 2021). For instance, Petrović et al. (2018) do that by assessing the variation of a scale-dependent segregation indicator. Their indicator measures the social diversity in each neighborhood's local environment, defined by a varying spatial extent around them, called scale. By computing the segregation indicator for a wide range of scales, and studying its maxima, they identify scales of interest. However, the indicator is aggregated at the city level, and does not convey the size, nor the number of segregated regions in the city.

Hitherto, studies have not drawn the size distribution of segregated regions in cities, while it could help understanding how segregation unfolds. This study proposes a direct approach to measure the size of segregated regions, by delineating their spatial extent. Using the proposed approach, we are able to make the following substantive contributions: we determine the size distribution of segregated regions per city, define the spatial scale of segregation for each city, and relate it to geographic, demographic, and urban characteristics of cities.

To delineate segregated regions in a city, we first determine the potential to encounter individuals from the social groups of interest in each spatial unit of the city — here defined as potential exposure —, and then aggregate these units into regions that are homogeneous with respect to this indicator. To measure the potential exposure to a group in a spatial unit, we compute the share of individuals from that group in the population able to walk to the spatial unit. The data required for this work consists in spatial demographics and street data. Agglomerative clustering is then used to aggregated the units into homogeneous regions. The regions in which the potential exposure is significantly larger than the city's average are labeled as segregated. Finally, one can compute the size of each segregated area, and define the scale of segregation from the size distribution of these areas.

In this study, we focus on the spatial segregation of people with a non-western migration background in the Netherlands. Segregation between the Dutch natives and the incoming population is now a major issue for local authorities, as it is deemed to hamper integration and social mobility (Zorlu and Mulder, 2008; Hartog and Zorlu, 2009; Vervoort, 2012; Tselios

et al., 2015). The spatial scale of segregation is particularly important in this context, as larger scale tends to be associated with lower integration (Tselios et al., 2015).

2 Data

This section summarizes the data sets used in the analysis. The two data sets used are demographic data (subsection 2.1), and data on the street layout (subsection 2.2). They are used to measure the potential exposure in each spatial unit. Demographic data set provides the share of each group residing in each spatial unit, and the street layout allows to determine the walking time between units.

2.1 Demographic data

Demographic data provides information on the population mix living in spatial units. It is retrieved from the Centraal Bureau voor de Statistiek (van Leeuwen, 2020). The spatial units are the 6-digits postcodes. These units are around 100x100 m^2 large, and populated by around 50 inhabitants in cities. Inhabitants are grouped into three categories: individuals from Dutch descent (both parents were born in the Netherlands), individuals with a western immigration background (Europe, North America, New Zealand, Australia, Japan), individuals with an non-western immigration background. In this data, individuals with a migration background are originating from another country than the Netherlands, or have one of their parents coming from another country than the Netherlands. Year 2017 is used for the analysis.

For privacy reason, the data is provided for a category in a spatial unit if at least 5 inhabitants belong to that category in that spatial unit, and if there are at least 10 residents living in the spatial unit. The data is provided in percentage terms of the total population in the spatial unit. The percentage is rounded to the closest 10%. We are able to partially correct these inaccuracies, using other variables in the data set (see appendix A for more details).

2.2 Street data

The street network is obtained from the OpenStreetMap (2021) data set, using the OSMnx library in Python (Boeing, 2017). This library allows to extract the street network in a given polygon (being the municipal border in our case). In order to compute the walking times between spatial units, we attach their centroids to the street network, and measure the walking distances between them.

3 Method

The objective of this work is to determine the scale of spatial segregation in a set of cities. For that, we first estimate the potential exposure to the different groups of interest in each spatial

unit (subsection 3.1), aggregate these units into homogeneous regions (subsection 3.2), and label these regions as segregated if the average exposure is significantly larger than the city average (subsection 3.3). Figure 1 below shows the different steps involved in the analysis. The map on the left depicts the share of the group of interest in the population residing in each spatial unit (using the raw data). The map in the center represents the exposure levels in each spatial unit, and the map on the right displays the regions designated as segregated. From this result, we can easily measure the size of each segregated region, and draw the size distribution for a given city. This section presents a situation with two social groups, but the method can easily be applied in a case study with more groups.

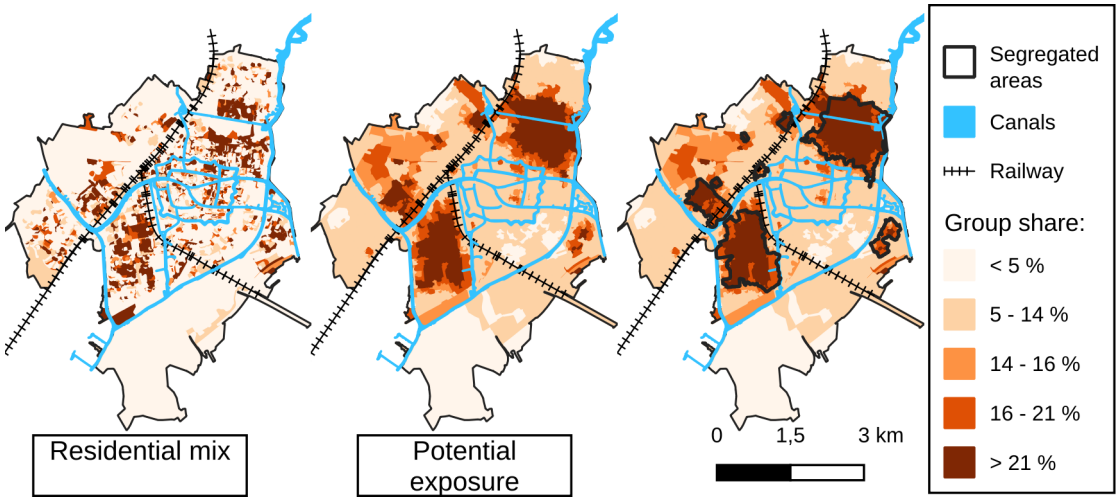


Figure 1: Delineation of segregated regions in Leiden. The color indicates the proportion of individuals from the group of interest among the population either residing in the spatial unit (left map), or able to reach it (center and right map).

3.1 Potential exposure

We compute the potential exposure in each spatial unit before aggregating them into homogeneous regions. The potential exposure to a given group in a spatial unit corresponds to the potential to encounter an individual from the group when walking in the spatial unit. This way, we identify regions that are homogeneous in terms of exposure. One could also aggregate units using the residential mix (left map in figure 1). However, the residential mix may represent poorly the segregation experienced by individuals: if one spatial unit highly populated by a group is surrounded by units deserted by that group, the segregation experienced by the inhabitants is most likely lower than what the residential mix indicates, as they still have decent opportunities to be exposed to different other. The potential exposure also has a smoother distribution (center map in figure 1), which is more suited for aggregating units into regions (Spierenburg et al., 2022).

We use an accessibility metric to quantify the ease with which people from each group can reach the centroid of the spatial unit. People able to reach the spatial unit are weighted using the walking time between their residence and the spatial unit: the further away people live,

the less likely they are to visit the spatial unit (subsection 3.1.1). Then, the potential exposure to a certain group in the spatial unit corresponds to the proportion of people from that group in the total number of people able to reach that spatial unit (subsection 3.1.2). This potential exposure is computed for all spatial units in all cities considered.

3.1.1 Travel impedance

The shortest walking distance from spatial unit to spatial unit is computed using the street network. The walking time is computed from the walking distance, using a walking speed of 4.5 km/h. Then, for a given destination spatial unit, we determine the origin spatial units located at an acceptable walking distance from it. We state that the walking time from a spatial unit to itself is 1 minute. All travel time shorter than 1 minutes are set to 1 minute. The inhabitants able to reach a given spatial unit are weighted using the walking time and the travel impedance function described by equation 1.

$$w(t) = \begin{cases} 1 & \text{if } 0 \leq t[s] < 60 \\ \frac{3600}{t^2} & \text{if } 60 \leq t < 1200 \\ 0 & \text{if } t \geq 1200 \end{cases} \tag{1}$$

The travel impedance shown in equation 1 is derived from the work of Schläpfer et al. (2021) providing a law to model visitation pattern of individuals in space. This law is expressed in equation 2. $\rho_i(r,f)$ is the influx of visitors coming to place i, living at a distance r from i and visiting i at a frequency f. μ_i is a constant depending on the place i, it relates to the attractiveness of i. η is 2 (derived empirically). The number of instances in which someone living at a distance r visits i at frequency f is $\rho_i(r,f) \cdot f$. Then the number of instances in which someone living at a distance r visits i is derived in equations 3 and 4. As we cannot estimate the attractiveness μ_i of a spatial unit i with the data we have, we assume it to be constant across all spatial units. We therefore have a simple function to model the number of visits of inhabitants from a spatial unit to another as a function of the distance separating the spatial units (equation 4). In the impedance function shown in 1, the distance r is replaced by the walking time t. We set a cut-off at a 20-minutes travel time to limit the number of shortest paths between spatial units to compute (we stop exploring a path if the length exceeds 20 minutes). The duration of travel times lasting less than 1 minute is not reliable (highly sensible to the location of the spatial units centroids), the impedance function is therefore set to be constant below 1 minute. The constant C is set to 3600 s^2, so that $w(t = 60s)$ is 1. This constant does not affect the potential exposure indicator (see next section).

$$\rho_i(r,f) = \frac{\mu_i}{(rf)^\eta} \tag{2}$$

$$w_i(r) = \int \rho_i(r,f) f \, df = \frac{\mu_i}{r^2} \int \frac{1}{f} df \tag{3}$$

$$w_i(r) = \frac{C}{r^2} \tag{4}$$

3.1.2 Potential exposure indicator

The likelihood that someone from group k living in spatial unit j visit spatial unit i is $N_{jk}\,w(t_{ij})$, where N_{jk} is the population living in j from group k. The likelihood that people from group k visit destination i is therefore the sum of likelihood over all origin spatial units (equation 5). Then the potential exposure to group k in spatial unit i is the share of visits from group k in all visits to spatial unit i. Constant C does not affect the potential exposure indicator: it is both in the numerator and the denominator of 6.

$$n_{ik} = \sum_j w(t_{ij}) \cdot N_{jk} \tag{5}$$

$$E_{ik} = n_{ik} / \sum_{k'} n_{ik'} \tag{6}$$

3.2 Detection of spatially segregated regions

After measuring the potential exposure in all spatial units, this work uses cluster analysis to group spatial units together into homogeneous regions in terms of exposure (subsections 3.2.1 to 3.2.3). Finally, regions in which the average exposure is significantly larger than the city average are labeled as segregated (subsection 3.3).

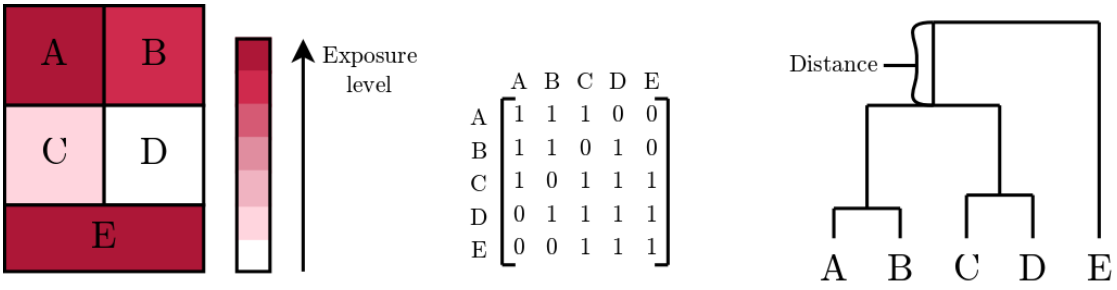
3.2.1 Agglomerative clustering

We use agglomerative clustering to group spatial units together into larger regions Theodoridis and Koutroumbas (2008). The variable of interest is the exposure to individuals with a non-western migration background. In the initialization phase, all spatial units are considered as independent clusters. Then, one merges clusters (hereafter called regions) iteratively. Merging regions that are not adjacent is forbidden (section 3.2.2). For each iteration, one considers the distance between each pair of regions, and merges the most similar regions together (in terms of potential exposure). The similarity between regions is determined by the Ward distance, shown in equation 7, minimizing the within-region variance (Müllner, 2011).

$$d(i \cup j, k) = \sqrt{\frac{(n_i + n_k)d(i,k) + (n_j + n_k)d(j,k) - n_k d(i,j)}{n_i + n_j + n_k}} \tag{7}$$

Figure 2 provides a toy example. At first, all spatial units (*A, B, C, D* and *E* on the left side of the figure) are considered as individual regions. The dissimilarity $d(i, j)$ between a pair (i, j) of these initial regions is the difference in their exposure level. The most similar regions are *A* and *E*, but they cannot be merged because they are not adjacent. Instead, *A* and *B* are merged. Then the distance between the newly formed region $A \cup B$ and every other region k is computed using formula 7, where n_i, n_j, and n_k are the number of initial spatial units in regions *A*, *B* and *C* respectively. We repeat the procedure until the stopping criteria is met

(subsection 3.2.3). One can also pursue the agglomeration until all spatial units are in the same region, allowing to build a dendrogram (see figure 2) summarizing all merging operations performed and the dissimilarity between all sub regions.



Map representation of exposure in a city composed of 5 zones.

Connectivity matrix corresponding to the city's topology.

Dendrogram obtained when aggregating zones together.

Figure 2: Example of agglomerative clustering applied to a toy city (left). The connectivity matrix (center) indicates the spatial units that are adjacent. The dendrogram (right) depicts the successive merging operations between regions.

3.2.2 Connectivity matrix

When aggregating spatial units together, one should ensure that spatial units are adjacent, so that regions are spatially continuous. For that, we use a connectivity matrix to ban merging operations that would result in spatially discontinuous regions. The connectivity matrix A for a city with N initial spatial units is a $N \times N$ matrix, in which component a_{ij} is one if i is adjacent to j, and zero otherwise (see center of figure 2).

3.2.3 Stopping criteria

The agglomerative process is stopped when the within-region variance exceeds a certain threshold. One need to find the optimal threshold. If the threshold is too high, the algorithm aggregates regions that do not have comparable exposure levels. If the threshold is too low, the algorithm misses aggregating regions that should belong to same larger region. We tune this parameter empirically by testing a wide range of values, and investigating the consistency of detected regions.

3.3 Labelling a region as segregated

The average exposure in a region R is computed from equation 8. In this equation, the contribution of spatial unit i is weighted by the population residing in the spatial unit. We label a region R as segregated if the average exposure in the region $\overline{y_R}$ is significantly larger than the average exposure μ in the city, when compared its standard deviation $\sigma_{\overline{y_R}}$ (see 9). In this equation, the region R is segregated when S is 1 (overexposure) or -1 (underexposure). When S is 0, the region is considered as mixed.

$$\overline{y_R} = \frac{\sum_{i \in R} n_i y_i}{\sum_{i' \in R} n_{i'}} = \sum_{i \in R} \theta_i y_i \tag{8}$$

$$S = \begin{cases} -1 & \text{if } \frac{\overline{y_R} - \mu}{\sigma_{\overline{y_R}}} \leq -1 \\ 0 & \text{if } -1 < \frac{\overline{y_R} - \mu}{\sigma_{\overline{y_R}}} < 1 \\ 1 & \text{if } \frac{\overline{y_R} - \mu}{\sigma_{\overline{y_R}}} \geq 1 \end{cases} \tag{9}$$

If two regions have the same value for S and are adjacent, they are merged.

To label a region as segregated, we need to compute the standard deviation of the average exposure in the region R, $\sigma_{\overline{y_R}}$. The derivation is included in appendix B.

4 Results

The aim of this work is to determine the scale of spatial segregation in all Dutch municipalities, and to assess how it associates with geographic, demographic, and urban characteristics. The method developed here is particularly suited to define the scale of segregation, as it draws the size distribution of segregated regions. The distribution provides different indicators (mean, median, largest component...) characterizing the scale of segregation. In this section, we first analyze the shape of the distribution, and determine a representative indicator for the scale (subsection 4.1). Then, we relate this scale indicator to city characteristics (subsection 4.2).

4.1 Size distribution of segregated regions, and scale of segregation in Dutch cities

With our method, we delineate around 800 segregated regions in all municipalities in the data set. Figure 3 shows how their size is distributed for the 3 largest cities: Amsterdam, Rotterdam and The Hague. The size is expressed in number of inhabitants, as the surfaces of regions are biased by large uninhabited areas (water, forest, parks...).

In most cities, the size of segregated regions spans over 1 to 2 orders of magnitude (figure 3 is in logarithmic scale). In Amsterdam for instance, the largest segregated region is populated by almost 190,000 inhabitants, and the smallest one is populated by around 5,000 inhabitants. The distribution is such that one or two regions are considerably larger than the rest. We use the median of the distribution as an indicator of the scale, so that half of the individuals living in a segregated region experience a scale larger or equal to the scale indicator. The median is preferred over the mean, as the distribution is skewed (few large regions, many small regions). In reality, for most cities, the median region is actually the largest one. In Amsterdam, the number of inhabitants living in a segregated region is around 370,000 inhabitants, and the largest component represents 51% of this number. Some cities like the Hague have only one large segregated component. Finally, the extent to which a region is segregated does not seem to correlate with its spatial scale. One could think that larger segregated regions are more

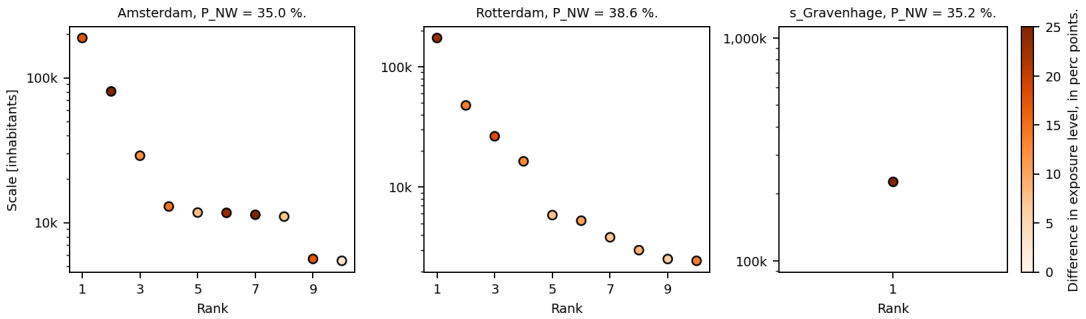


Figure 3: Rank-size distribution of segregated regions in the 3 largest Dutch municipalities. In each municipality, segregated regions are ordered by population (in thousands inhabitants) and plotted against the region's rank. The scale is logarithmic. The share of individuals with a non-western migration background living in the city is expressed by the variable *P_NW* next to the city name. The color of the point represents how much larger the exposure to individuals with a non-western migration background is compared to the city average (using the difference).

mixed, but this is not observed in figure 3. In this figure, the color represents how much a region is overexposed to a certain group compared to the city average. Larger regions do not seem to be either more or less overexposed than smaller ones. This is also the case for all other cities considered in the analysis.

4.2 Association of the scale with city characteristics

Now that the scale of spatial segregation is clearly defined for each city, we investigate how the scale relates to geographic, demographic, economic and urban characteristics of cities. In the literature, the characteristics usually correlated with segregation level are the city size, income inequality, fragmentation of space by physical boundaries, urban sprawl, and the scale and pace at which affordable housing has been developed (Gordon and Monastiriotis, 2006; Natale et al., 2018; Ananat, 2011; Andersen et al., 2016; Hess et al., 2021). In this work, instead of relating these characteristics to the segregation level in cities, we relate them to the scale of segregation. Here, the scale of segregation is associated to the city size (subsection 4.2.1), economic and demographic variables (subsection 4.2.2), and characteristics of the urban environment (subsection 4.2.3).

4.2.1 Relation between the scale of segregation and the city size

We observe a striking correlation between the scale of segregation (represented by the number of inhabitants living in the median segregated region) and the population in the city (see figure 4). The correlation coefficient is 0.92. One could expect some correlation as the indicator for the scale is constrained by the city population: the scale cannot be larger than the total population. Such dependence would usually result in correlation, regardless of the segregation pattern. Then, one can expect two different patterns: the relation between the scale of segregation and the city population could be either linear, or sub-linear. In the linear situation, larger cities would tend to have larger segregated regions than medium cities, while in the

sub-linear situation, they would tend to have more medium-sized segregated regions (given equal levels of segregation). Figure 4 shows a clear linear relationship between the scale of segregation and the city size rather than a sub-linear relationship. The scale of segregation in a city with twice more inhabitants than another is expected to be twice larger than in that other city.

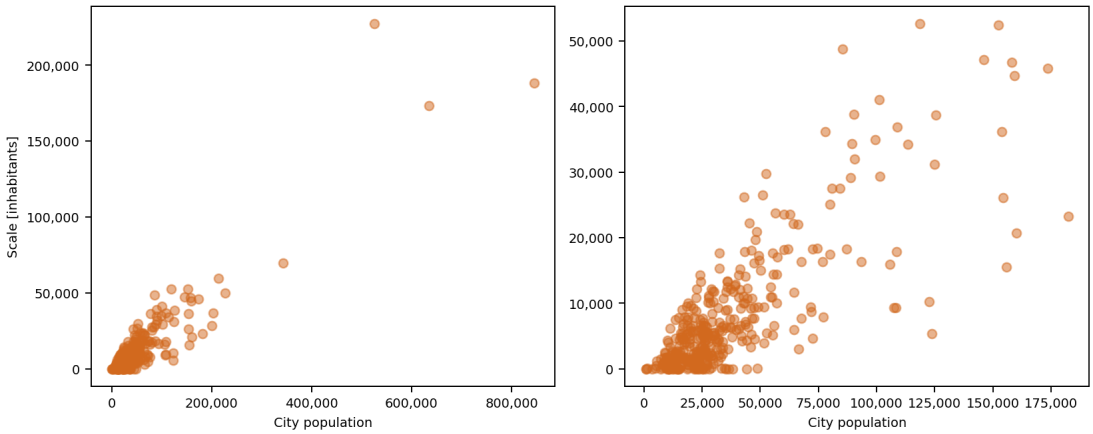


Figure 4: Relation between the scale of segregation and the city population. The plot on the left shows all municipalities in the Netherlands, while the one on the right filters out the ones with more than 200,000 inhabitants.

4.2.2 Relation between the scale of segregation and demographic and economic characteristics

In cities, the segregation level is usually associated with demographic and economic characteristics. Segregation is larger in cities when the share of groups is more even (the minorities shares are close to the majority share), and larger for more affluent and unequal cities (OECD, 2018). In this section, we determine whether these characteristics also correlate with the scale of segregation. Figure 5 shows the relation between scale and income inequality (represented by the Gini coefficient on the left plot), and the share of the group of interest in the total population of the city (right plot). The size of the dot in these plots is proportional to the city population in the city. In these two figures, the scale is represented in relative terms (population in the median segregated area divided by the city population).

We observe that cities in which the group of interest represents a larger share of the population show larger-scale segregation (correlation coefficient of 0.34). This relation seems to be sub-linear, the shape of the scatter plot is concave. This could be explained by the fact that, on one hand, segregated regions take more space as more people belong to the group of interest (increasing the scale); and on the other hand, the group of interest represents a larger share of the population in segregated regions, (limiting the increase in the scale).

Income inequality does not seem to be related to the scale of segregation, the correlation coefficient between the scale relative to the city population and the Gini coefficient is 0.03.

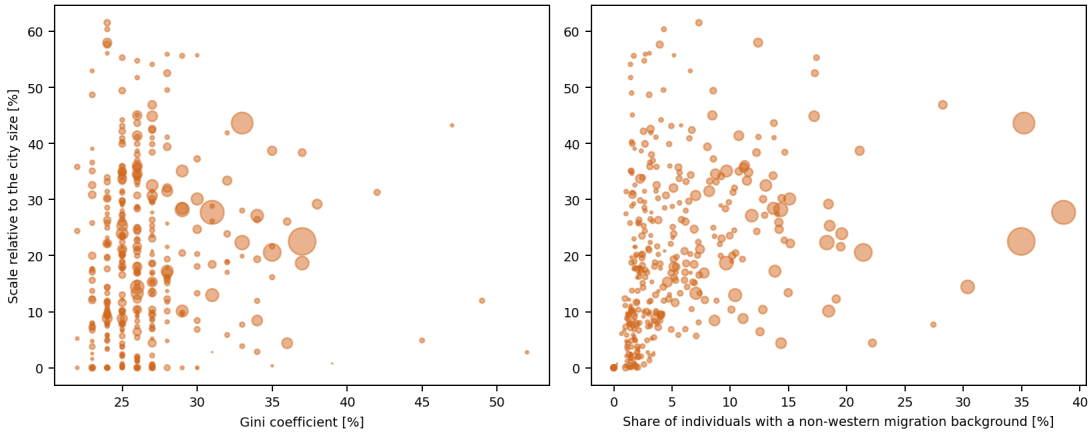


Figure 5: Relation between the scale of segregation and economic and demographic characteristics of cities. The plot on the left shows the relationship between the scale of segregation and the Gini coefficient. The plot on the right shows the relationship between the social composition of the city and the scale of segregation. The size of the dots is proportional to the city population.

4.2.3 Relation between the scale of segregation and urban characteristics

Features of urban development are also deemed to be associated with certain segregation levels. Studies have shown that cities that are denser, more divided by physical boundaries (as railways), and that have experienced large-scale affordable housing development in their history tend to be more segregated (Ananat, 2011; Hess et al., 2021). This subsection relates indicators on these three urban components to the scale of segregation. Figure 6 displays the relation observed in the Netherlands.

The density of the city is measured using the population to density allocation, being the share of the population living in spatial unit that is denser than a certain density threshold (set to 3500 inhabitants per square kilometer in this case). A population to density allocation of 0.5 means that 50% of the individuals live in a spatial unit that is denser than 3500 inhabitants per square kilometer. We prefer this indicator over the density, as it is not biased by potential uninhabited areas inside the city (river, canals, forest...), and relates more to the density experienced by inhabitants. The left plot in figure 6 shows that denser cities experience larger scale segregation (correlation coefficient of 0.42).

By large-scale affordable housing development we mean the waves of high-rise buildings built in the 60s-70s in the Netherlands to address an intense demand in affordable housing in cities. Cities in which such housing estate were developed are deemed to be more segregated (Hess et al., 2021). To investigate this relation, we propose the housing entropy index (equation 10), that measures how consistent are spatial units with respect to housing development. Spatial units in which most houses have been developed in the same time period are considered as more consistent than units mixing houses developed in different time periods. As an indicator we use the entropy index often used in segregation studies to determine how uniform/mixed spatial unit are in terms of demographics (Reardon and O'Sullivan, 2004). In this case, we investigate how uniform/mixed are spatial units in terms of housing development period. For

each spatial unit, the demographic data also provides the number of housing built within a certain time window (before 1945, between 1945 and 1964...). We can then use the entropy index (equation 10) to determine the extent to which spatial units mix housing development periods. In this equation, p_{gi} represents the share of housing p_{gi} from period g in the total number of housing h_i in spatial unit i, and $\overline{p_g}$ represents the share of housing built in period g in the total number of housing in the city. This indicator H is comprised between 0 and 1, 1 being a situation in which spatial units are only composed by housing developed in one period. We observe a weak positive correlation between this indicator and the scale of segregation (correlation coefficient of 0.28).

$$H = 1 - \frac{\sum_i \sum_g h_i p_{gi} log(p_{gi})}{\sum_g \overline{p_g} log(\overline{p_g})} \tag{10}$$

The division of space is measured using the division index proposed by Ananat (2011), see equation 11. This indicator represents the extent to which space is divided by physical boundaries, comprised between 0 and 1 (the larger, the more divided). In this indicator, a is the index of the fragments of the city divided by physical boundaries, and S_a is the surface of fragment a. Many types of physical boundaries could be considered (railways, highways, canals...), in this work, we choose to consider only railways. On figure 6, we do not identify a clear relation between the division of space and the scale of spatial segregation, and the correlation is weak (correlation coefficient of -0.23).

$$DI = 1 - \sum_a \frac{S_a^2}{S_{tot}^2} \tag{11}$$

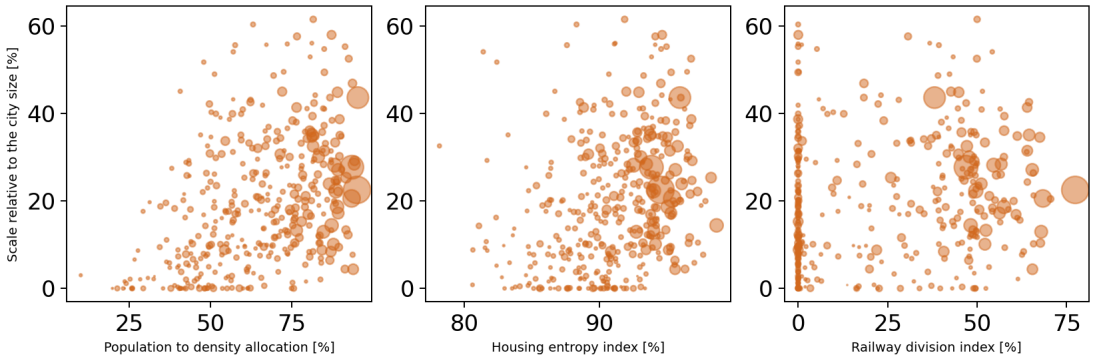


Figure 6: Relation between the scale of segregation and indicators characterizing urban development in cities. The size of dots is proportional to the city population.

5 Conclusion and outlook

This study proposes a novel data-driven approach to delineate the geographical demarcation of segregated regions. This approach allows to measure directly the spatial scale of segregation

from the size distribution of segregated regions, while previous studies would investigate the scale indirectly, by assessing how a segregation index varies with a scale parameter. We have then investigated how the scale relates to city characteristics. We observe a striking correlation of the scale of segregation with the city size, and moderate correlation with income inequality, share of the group of interest in the city, and population density.

Further research will be devoted to determine these relations at the segregated region level, instead of the city level. The analysis will be replicated at different timestamp, allowing to study the evolution of segregated regions in time, and for other countries, to assess how the scale of segregation relates to the migration history and the housing market in countries.

References

Ananat, E. O. (2011). The wrong side(s) of the tracks: The causal effects of racial segregation on urban poverty and inequality. *American Economic Journal: Applied Economics*, 3:34–66.

Andersen, H. S., Andersson, R., Wessel, T., and Vilkama, K. (2016). The impact of housing policies and housing markets on ethnic spatial segregation: comparing the capital cities of four nordic welfare states. *International Journal of Housing Policy*, 16(1):1–30.

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

Gordon, I. and Monastiriotis, V. (2006). Urban size, spatial segregation and inequality in educational outcomes. *Urban Studies*, 43:213–236.

Hartog, J. and Zorlu, A. (2009). Part 1: Ethnicity, religion, discrimination and exclusion: Ethnic segregation in the netherlands: An analysis at neighbourhood level. *International Journal of Manpower*, 30:15–25.

Hess, D. B., Tammaru, T., and van Ham, M. (2021). *The Urban Book Series Housing Estates in Europe Poverty, Ethnic Segregation and Policy Challenges*, chapter Lessons Learned from a Pan-European Study of Large Housing Estates: Origin, Trajectories of Change and Future Prospects. Springer.

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R., and Sanbonmatsu, L. (2012). Neighborhood effects on the long-term well-being of low-income adults. *Science*, 337.

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv*.

Natale, F., Scipioni, M., and Alessandrini, A. (2018). *Divided Cities: Understanding intra-urban inequalities*, chapter Spatial segregation of migrants in EU cities. OECD Publishing.

OECD (2018). *Divided Cities*. OECD Publishing.

OpenStreetMap (2021). OpenStreetMap data. Last accessed in June 2021, more info at https://www.openstreetmap.org/.

Owens, A. (2018). Income segregation between school districts and inequality in students' achievement. *Sociology of Education*, 91:1–27.

Petrović, A., van Ham, M., and Manley, D. (2018). Multiscale measures of population: Within- and between-city variation in exposure to the sociospatial context. *Annals of the American Association of Geographers*, 108(4):1057–1074.

Reardon, S. F., Matthews, S. A., O'sullivan, D., Lee, B. A., Firebaugh, G., Farrell, C. R., and Bischoff, K. (2008). The geographic scale of metropolitan racial segregation. *Demography*, 45:489–514.

Reardon, S. F. and O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology*, 34:121–162.

Schläpfer, M., Dong, L., O'Keeffe, K., Santi, P., Szell, M., Salat, H., Anklesaria, S., Vazifeh, M., Ratti, C., and West, G. B. (2021). The universal visitation law of human mobility. *Nature*, 593:522–527.

Spierenburg, L., van Cranenburgh, S., and Cats, O. (2022). A regionalization method filtering out small-scale spatial fluctuations. *AGILE: GIScience Series*, 3:61.

Tammaru, T., van Ham, M., Marcińczak, S., and Musterd, S. (2015). *Socio-Economic Segregation in European Capital Cities: East Meets West*. Routledge.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition*, chapter 13: Clustering Algorithms II: Hierarchical Algorithms, pages 654–658. Academic Press, fourth edition.

Tselios, V., Noback, I., van Dijk, J., and McCann, P. (2015). Integration of immigrants, bridging social capital, ethnicity, and locality. *Journal of Regional Science*, 55(3):416–441.

Tóth, G., Wachs, J., Clemente, R. D., Ákos Jakobi, Ságvári, B., Kertész, J., and Lengyel, B. (2021). Inequality is rising where social network segregation interacts with urban topology. *Nature Communication*, 12.

van Leeuwen, N. (2020). *Statistische gegevens per vierkant en postcode 2019–2018–2017*. Centraal Bureau voor de Statistiek. Retrieved from https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode in June 2021.

Veneri, P., Comandon, A., Àngel Garcia-López, M., and Daams, M. N. (2021). What do divided cities have in common? an international comparison of income segregation. *Journal of Regional Science*, 61:162–188.

Vervoort, M. (2012). Ethnic concentration in the neighbourhood and ethnic minorities' social integration: Weak and strong social ties examined. *Urban Studies*, 49:897–915.

Wang, Q., Phillips, N. E., Small, M. L., and Sampson, R. J. (2018). Urban mobility and neighborhood isolation in america's 50 largest cities. *Proceedings of the National Academy of Sciences of the United States of America*, 115:7735–7740.

Wilf, H. S. (2005). *generatingfunctionology*. Academic Press, Inc.

Williams, D. R. and Collins, C. (2001). Racial residential segregation:a fundamental cause of racial disparities in health. *Public Health Reports*, 116:404–416.

Wilson, W. (1987). *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Sociology, urban studies, black studies. University of Chicago Press.

Zorlu, A. and Mulder, C. H. (2008). Initial and subsequent location choices of immigrants to the netherlands. *Regional Studies*, 42:245–264.

A Preprocessing of demographic data

The values per zone in the data is rounded, and this decreases the accuracy of the results. For the population, if a zone is composed of 8 inhabitants, the rounded value will be 10, representing a 25% error. For the proportion of individuals with a migration background, if individuals with a non-western migration background represent 5% of the population in the zone, it will be rounded to 10%, which corresponds to a 100% error. This section describes the method used to make an estimate of the demographics that is closer than the rounded values provided.

A.1 Estimation of the population from the age groups and gender

The population living in a zone can be estimated using the population per gender, per age group, and the number of households. For instance, if the dataset indicates a population of 15 individuals, 10 women and 10 men, the actual population is either 16 (8 women and 8 men) or 17 (9 women, 8 men). The population cannot be larger than 17 (otherwise rounded to 20), or smaller than 17 (otherwise the population per gender is incorrect). The same holds for the 5 age groups. The paragraph below describes the method used using the age groups as an example.

One must differentiate three cases:

- The population in an age group is strictly lower than 5. In this case, the data will not provide any value (*None*). Therefore, if the data is *None* for an age group, the population in this age group is 0, 1, 2, 3 or 4. In this case, the *None* value is replaced by 2, corresponding to the median.
- The population in an age group is comprised between 5 and 7. In this case, the population is rounded to the closest 5. The Therefore, if the data is 5, the population is either 5, 6 or 7 (if smaller than 5, it falls in the previous category).
- The population in an age group is strictly larger than 7. In this case, the population is rounded to the closest 5. In such case, the population can be the rounded population, the rounded population minus 2, minus 1, plus 1, or plus 2.

The value for the rounded population $\tilde{A}_i$ of an age group i in the data is either *None*, 5, or a multiple of 5 (see table 1). The true population A_i of that age group is a random variable that can take the values specified in table 1. The difference D_i between A_i and the rounded value $\tilde{A}_i$ is also a random variable. The total population N is equal to the sum A of the random variables A_i. The difference D between the actual total population N and the sum of all rounded age groups $\sum_i \tilde{A}_i$ is the sum of all D_i. Therefore, the probability that the actual population in a zone is $\tilde{N} + d$ given the population in all age groups is $P(D = d \mid \tilde{A})$. This probability is the number of combinations $(D_1, D_2, D_3, D_4, D_5)$ summing to d among all possible combinations of $(D_1, D_2, D_3, D_4, D_5)$. This is done using the concept of generating functions (Wilf, 2005). The generating function for variable D_i is provided in the last column of table 1.

$\tilde{A}_i$	Possible values for A_i	Ref. value	Possible values for D_i	Generative function
None	(0,1,2,3,4)	2	(-2,-1,0,1,2)	$x^{-2}+x^{-1}+x^0+x^1+x^2$
5	(5,6,7)	5	(0,1,2)	$x^0+x^1+x^2$
>5	($\tilde{A}_i$-2,$\tilde{A}_i$ -1,$\tilde{A}_i$,$\tilde{A}_i$ +1,$\tilde{A}_i$ +2)	$\tilde{A}_i$	(-2,-1,0,1,2)	$x^{-2}+x^{-1}+x^0+x^1+x^2$

Table 1: Generative function corresponding to the random variable D_i.

$$D = D_1 + D_2 + D_3 + D_4 + D_5 \tag{12}$$
$$g_{N\neq 5}(x) = (x^{-2} + x^{-1} + x^0 + x^1 + x^2) \tag{13}$$
$$g_{N=5}(x) = (x^0 + x^1 + x^2) \tag{14}$$
$$g(x) = g^{\lambda}_{N\neq 5}(x) \cdot g^{(5-\lambda)}_{N=5}(x) \tag{15}$$
$$g(x) = x^{-2\lambda}(x^0 + x^1 + x^2 + x^3 + x^4)(x^0 + x^1 + x^2)^{5-\lambda} \tag{16}$$

The following derivations determine the coefficient of x^d in the generative function of D.

$$g(x) = x^{-2\lambda}\left(\frac{1-x^5}{1-x}\right)^{\lambda}\left(\frac{1-x^3}{1-x}\right)^{5-\lambda} \tag{17}$$
$$g(x) = x^{-2\lambda}(1-x^5)^{\lambda}(1-x^3)^{5-\lambda}(1-x)^{-5} \tag{18}$$

Equation 18 can be written as a power series using equation 24. The negative binomial coefficient in equation 19 is computed using equation 25.

$$g(x) = x^{-2\lambda}\sum_{k=0}^{\lambda}\binom{\lambda}{k}(-1)^k x^{5k}\sum_{i=0}^{5-\lambda}\binom{5-\lambda}{i}(-1)^i x^{3i}\sum_{j=0}^{\infty}\binom{-5}{j}(-1)^j x^j \tag{19}$$
$$g(x) = x^{-2\lambda}\sum_{k=0}^{\lambda}\binom{\lambda}{k}(-1)^k x^{5k}\sum_{i=0}^{5-\lambda}\binom{5-\lambda}{i}(-1)^i x^{3i}\sum_{j=0}^{\infty}\binom{j+4}{4}x^j \tag{20}$$

The exponent d of a term x^d in the generative function of d in equation 20 can be expressed as a function of λ, k, i and j (equation 21). Hence, j in equation 20 can be replaced by j in equation 22.

$$d = -2\lambda + 5k + 3i + j \tag{21}$$
$$j = d + 2\lambda - 5k - 3i \tag{22}$$

Then, the coefficient of x^d in equation 20 (noted $[x^d]g(x)$) is given by equation 23 below, corresponding to the number of combination of $(D_1, D_2, D_3, D_4, D_5)$ for which the D_i sum up to d.

$$[x^d]g(x) = \sum_{k=0}^{\lambda} (-1)^k \binom{\lambda}{k} \sum_{i=0}^{5-\lambda} (-1)^i \binom{5-\lambda}{i} \binom{d+2\lambda-5k-3i+4}{4} \tag{23}$$

$$(x+y)^n = \sum_{k=0}^{\infty} \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{24}$$

$$\binom{-5}{j} = (-1)^j \binom{5+j-1}{j} = (-1)^j \binom{j+4}{4} \tag{25}$$

$$\binom{n}{k} = \binom{n}{n-k} \tag{26}$$

The probability that the sum of all D_i is d is computed in equation 27 below.

$$P(D = d \mid \tilde{A}) = \frac{[x^d]g(x)}{5^\lambda \cdot 3^{(5-\lambda)}} \tag{27}$$

The raw data also provides the rounded total population in the zone $\tilde{N}$, constraining the possible values of N (equation 28), and d (equation 29).

$$\tilde{N} - 2 \leq N \leq \tilde{N} + 2 \tag{28}$$

$$\tilde{N} - \sum_i \tilde{A}_i - 2 \leq d \leq \tilde{N} - \sum_i \tilde{A}_i + 2 \tag{29}$$

Finally, the probability that the actual population N is n given $\tilde{N}$ and all $\tilde{A}_i$ is given in equations 30 and 31.

$$P(N = n \mid \tilde{N}, \tilde{A}) = P(D = d \mid \tilde{N}, \tilde{A}) \tag{30}$$

$$= \frac{P(D = d \mid \tilde{A})}{\sum_{d'} P(D = d' \mid \tilde{A})}, \text{ such that } \tilde{N} - \sum_i \tilde{A}_i - 2 \leq d' \leq \tilde{N} - \sum_i \tilde{A}_i + 2 \tag{31}$$

The same approach is used to estimate the population using the gender categories G_i, and the number of households using the categories of households.

A.2 Estimation of the population from the data on households

One can also use the number of households and the average number of people per household to estimate the population in the zone. If the number of households is 10 in the data (the

true value is comprised between 8 and 12), and there are 2.1 individuals per households on average, one can evaluate the combination (n,h), with h the number of households, that would result in the correct number of individuals per household η_h. One should remember that the number of households $\tilde{H}$ in the raw data is rounded to the closest 5 and η_h is rounded to the first decimal. We first explore the set S of possible combinations (n,h), given the following conditions:

$$n \in \{\tilde{N}-2, \tilde{N}-1, \tilde{N}, \tilde{N}+1, \tilde{N}+2\} \tag{32}$$

$$h \in \{\tilde{H}-2, \tilde{H}-1, \tilde{H}, \tilde{H}+1, \tilde{H}+2\} \tag{33}$$

$$\left[\frac{n}{h}\right]_{\text{rounded to 0.1}} = \eta_h \tag{34}$$

The probability that the actual population N is n given $\tilde{H}$ is computed by counting the combination (n,h) among all combinations (N,h) in S. The combinations are weighted using the probability $P(H = h \mid \tilde{H}, \tilde{C})$ that the actual number of household is h given $\tilde{H}$ and the number of household $\tilde{C}_i$ per category i. This probability is computed using the method described in subsection **??**, where the 5 age groups are replaced by the 4 categories of households.

$$P(N = n \mid \tilde{H}, \eta_h, \tilde{C}) = \frac{\sum_h \delta_{(n,h)} P(H = h \mid \tilde{H}, \tilde{C})}{\sum_{n'} \sum_{h'} \delta_{(n',h')} P(H = h' \mid \tilde{H}, \tilde{C})} \qquad \delta_{(n,h)} = \begin{cases} 1 & \text{if } (n,h) \in S \\ 0 & \text{if } (n,h) \notin S \end{cases} \tag{35}$$

A.3 Estimation of the population

The estimation of the population in a zone is determined by combining the estimation of the population using the age groups $P(N = \tilde{N} + d|\tilde{A})$, the gender categories $P(N = \tilde{N} + d|\tilde{G})$, and the data on households $P(N = \tilde{N} + d|\tilde{H})$ (equation 36).

$$P(N = \tilde{N} + d) = \frac{P(N = \tilde{N} + d|\tilde{A}) \cdot P(N = \tilde{N} + d|\tilde{G}) \cdot P(N = \tilde{N} + d|\tilde{H})}{\sum_{d'} P(N = \tilde{N} + d'|\tilde{A}) \cdot P(N = \tilde{N} + d'|\tilde{G}) \cdot (N = \tilde{N} + d'|\tilde{H})} \tag{36}$$

The population estimate $\hat{N}$ is defined from the expectation of N, rounded to the closest integer.

B Computing the variance of the average exposure in a region

The variance of the average exposure in a region $\overline{y_R}$ can be computed analytically from the equations below.

$$\overline{y_R} = \sum_{i \in R} \theta_i y_i \tag{37}$$

$$Var(\overline{y_R}) = \sum_{i \in R} \sum_{j \in R} Cov(\theta_i y_i, \theta_j y_j) \tag{38}$$

$$= \sum_{i \in R} \sum_{j \in R} \theta_i \theta_j Cov(y_i, y_j) \tag{39}$$

The coefficients θ_i are computed from equation 40.

$$\theta_i = \frac{n_i}{\sum_{i' \in R} n_{i'}} \tag{40}$$

The covariance matrix Σ containing the covariances $Cov(y_i, y_j)$ can be derived analytically, from the definition of y_i (see equation 41). In this equation, $w(t_{ik})$ is the travel impedance when walking from k to zones i, and x_k is the share of individual from the group of interest residing in zone k. The derivation of the covariance coefficients can be found in equations 43 to 45.

$$y_i = \frac{\sum_k w(t_{ik}) \cdot n_k \cdot x_k}{\sum_{k'} w(t_{ik'}) \cdot n_i} \tag{41}$$

$$y_i = \sum_k c_{ik} x_k \tag{42}$$

$$Cov(y_i, y_j) = Cov\left(\left(\sum_k c_{ik} x_k\right), \left(\sum_{k'} c_{jk'} x_{k'}\right)\right) \tag{43}$$

$$= \sum_k \sum_{k'} c_{ik} \cdot c_{jk'} \cdot Cov(x_k, x_{k'}) \tag{44}$$

$$= \sum_k c_{ik} \cdot c_{jk} \cdot Var(x_k) \tag{45}$$

In equation 44, the covariance of the two variables x_k and $x_{k'}$ is given in equation 46. To assess the significance of the average $\overline{y_R}$ in relation to μ in equation 9, we compare to a case where the variables x_k are randomly allocated to the zone k (the population n_k and the travel impedance $w(t_{ik})$ remain the same). This allows to assess the extend to which the average $\overline{y_R}$ can be observed by luck. The variables x_k are randomly allocated, they are therefore independent. The covariance $Cov(x_k, x_{k'})$ is 0 when $k \neq k'$. The variance $Var(x_k)$ is obtained from the actual distribution of x (the share of individuals from the group of interest residing in a zone).

$$Cov(x_k, x_{k'}) = \begin{cases} 0 & \text{if } k \neq k' \quad \text{(Uncorrelated variables)} \\ Var(k) & \text{if } k = k' \end{cases} \tag{46}$$

The value of coefficient c_{ik} is expressed in equation 47.

$$c_{ik} = \frac{w(t_{ik}) \cdot n_k}{\sum_{k'} w(t_{ik'}) \cdot n_{k'}} \tag{47}$$

This computation can be summarized by the matrix multiplication shown in equation 48, where C is given in equation 49.

$$\Sigma = C^T \times C \cdot \sigma_x^2 \tag{48}$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1N} \\ c_{21} & & & \\ \vdots & & & \\ c_{N1} & & & c_{NN} \end{bmatrix} \tag{49}$$