Volume 12, Number 1, 2025

REGION

The Journal of ERSA
Powered by WU

Table of Contents

Articles

Funded by

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 1, 2025, 1–18
DOI: 10.18335/region.v12i1.521

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

Spillover Effects of Public Capital Stock: A Case Study for Ecuador

Roberto Zurita[1], Víctor Morales-Oñate[2]

[1] Universidad San Francisco de Quito, Quito, Ecuador
[2] Universidad de las Américas, Quito, Ecuador

Received: 5 November 2023/Accepted: 13 January 2025

Abstract. This research examines the spatial spillovers of public capital on gross value added across 216 cantons in continental Ecuador. The investigation is conducted within the framework of Spatial Econometrics, utilizing various model specifications and spatial weight matrices, complemented by a Cobb Douglas-type model that incorporates spatial dependence. The findings highlight a positive spatial impact of the public capital stock, with approximately 30% of the overall effect attributed to the indirect component. This underscores the importance of considering spatial structure when assessing the effects of capital on gross value added. Consequently, the study extends its exploration to derive column and row effects, aimed at identifying the most influential cantons within the neighborhoods established by the spatial structure.

1 Introduction

The literature has found evidence of the effects of public capital on the economic performance of countries, as it is a factor that, together with private capital, labor and technology, contributes to productive performance. However, new research in the field of new economic geography has revealed that these factors can spread their effect in spatial dimensions to nearby territories.

Spatial economics seeks to explain the causes of the unequal distribution of wealth among territories, understanding what factors attract and concentrate economic activities to a site and what forces cause their dispersion. This field of study allows us to incorporate the conceptual framework of economics in the spatial dimension, in order to understand economic phenomena at the regional level (Marrocu, Paci 2010).

According to economic orthodoxy, it is known that the level of production of a country depends on factors such as labor, capital stock and technology. In relation to the capital that a country possesses, we understand the means of production that companies possess to produce, but we must also take into account the stock of public capital, since it has been inferred that investment in public infrastructure such as roads, railroads, basic services, among others, contribute to lowering the production costs of companies and allow mobilizing labor to production centers, thus improving the economic performance of the regions (Fingleton 2001).

Most developed countries present spatial economic structures that base value generation on research, innovation, high-tech industry, service provision, etc. On the other hand, developing countries largely maintain a productive matrix based on the generation of primary goods linked to natural resources, and with a low level of productive linkages.

This has influenced the way in which their economic activities have been distributed throughout their territories, and the way in which they interact with each other. Within this dynamic, it is of special interest to analyze how production factors such as public capital influence economic performance in the territories of a developing country, and how their effect can spread to nearby spatial units.

The literature has found evidence of the effects of public capital on the economic performance of countries, being a factor that together with private capital, labor and technology contribute to productive performance. However, new research in the field of new geographical economics and spatial economics such as Han et al. (2016) have revealed that these factors can spread their effect in spatial dimensions to nearby territory.

The majority of research defines public capital as the physical assets owned by the government, excluding military-related assets (Bom, Ligthart 2014). This implies that both public and private capital play a role in creating a conducive economic environment. Consequently, there has been significant scholarly endeavor aimed at quantifying the impact of public capital on economic performance.

Mera (1973) stands as one of the initial contributors to the field, delving into the impacts of public capital. Employing econometric techniques with both additive and multiplicative production functions, this study utilized ordinary least squares. Notably, Mera's research unearthed early signs that the influence of production elasticity concerning public capital heavily relies on how this variable is defined. Notably, elasticities demonstrated notably higher values when encompassing transportation infrastructure. The study was conducted across 46 Japanese prefectures during the span of 1954 to 1963. Although evidence has been found that public capital improves the economic performance of regions, it is necessary to categorize it. Not all public investment has a significant influence on firm productivity.

Bom, Ligthart (2014) categorize public capital into two groups: i) Central or core, which includes highly productive infrastructure like roads, railways, and airports, as well as key public services such as sewage and water systems due to their direct impact on economic activity, and ii) Non-central or peripheral, which encompasses other public services and structures, including hospitals, educational facilities, and various other public buildings. Aschauer (1989) delves into the distinct impacts of core and non-core public capital. Employing the production function, he sought to understand the decline in productivity growth in the US during the 1970s. He discovered that a 1% rise in the core public capital stock led to a 0.39% boost in private production. This significant figure indicates that public capital played a pivotal role in influencing production.

Berndt, Hansson (1992) concentrate exclusively on the role of core capital in enhancing the private sector's productivity performance. They investigated how it reduced production costs within the Swedish economy during the 1980s. One of their significant findings was that core public infrastructure played a pivotal role in cost reduction for the private sector. Through counterfactual simulations, they demonstrated that the Swedish economy could have mitigated its productivity slowdown by 6.1% if it had adhered to optimal public spending levels. In doing so, the authors identified a mechanism through which public investment could enhance the productivity of the private sector.

Since that time, many studies have been conducted for the United States as well as several OECD nations. More recently, the impact of public capital on productivity in developing countries has also garnered attention. Ram (1996) examined the roles of both public and private capital in these countries throughout the 70s and 80s. His findings suggest that during the 70s, private capital outperformed public capital in terms of productivity. However, in the 80s, public capital took the lead, contributing more to production than private capital.

Guevara (2016) demonstrates through spatial econometric methods that urban agglomerations generate a spatial spillover effect of their economic growth to their neighboring regions in Latin American countries such as Argentina, Bolivia, Chile, Colombia, Ecuador, Mexico, Peru and Panama. However, the sample used does not include all the regions of the countries analyzed. Álvarez et al. (2016) perform a spatial econometric analysis to determine the spillover effects of public capital as a factor of the production function for the regions of Spain for the periods 1980-2007. The findings show that

transportation infrastructure generates a positive and significant spillover effect across regions. Jia et al. (2020) conduct a spatial analysis between factors of different production functions among rural regions in Taizhou municipality in China, finding evidence of spatial correlation between regions with different production patterns.

In the context of Ecuador, research has been conducted to evaluate the elasticities of GDP in relation to production factors like capital and labor. Briones Bendoza et al. (2018) undertook an analysis of the variations in these factors from 1950 to 2014. They employed an econometric approach, leveraging ordinary least squares. Their findings suggest that physical capital plays a more significant role in production compared to labor. This trend might be attributed to the nation's dominant economic activities relying on low-skilled, low-wage labor, thus amplifying the relative contribution of capital. However, this study does not distinguish between public and private capital, making it challenging to discern the specific contributions of each. Moreover, the study's capital variable represents gross capital, encompassing both private and public capital, including its core, non-core, and military segments. In light of this, as per Bom, Ligthart (2014) and Aschauer (1989), the non-core capital likely has limited influence on production, and military expenditure is anticipated to be non-influential.

Moreno Loza (2017) delves into the implications of fiscal policy in Ecuador between 2000 and 2015, aiming to assess the impact of current spending, capital spending and tax revenue on gross domestic product (GDP). This investigation employs the VARS (structural vector autoregressive) model for the analysis. The predominant findings indicate that fiscal modifications directed towards capital expenditure yield a multiplier effect of 0.37 on GDP, marking it as the most influential category. Conversely, alterations in current public expenditure yield a multiplier effect of 0.11 on GDP. It is worth noting that this study primarily focuses on a national scope, without exploring the resultant effects on economic performance or the productivity discrepancies across different regions.

Most of the cited literature on the evidence of economic spillover effects from public capital has been conducted in industrialized countries with higher levels of public capital stocks compared to those in a developing economy such as Ecuador. In a spatial econometrics setting, growth within a specific region is determined by the independent variables across all other regions within the system. This is the mechanism by which public capital in one canton can influence on the economic growth of neighbors. Therefore, this research contributes to finding evidence of contagion effects in a developing economy and understanding how these economic effects are transmitted among its regions. Indirect effects are spillover effects and direct effects include feedback effects. Spatial dependence structure is examined by setting different weights matrices. Finally, differentiating private and public capital across 216 cantons in continental Ecuador implied a detailed information gathering exercise which let us apply spatial models.

This paper seeks evidence that a production factor such as public capital can generate spatial effects on the production levels of the different cantons of Ecuador. For this purpose, a spatial econometric analysis is carried out through different types of models that allow sensitizing the economic analysis of the production factors with the geographic structures that can generate effects on the economic dynamics of a nation. To this end, data were collected for the year 2017 from 216 cantons nationwide.

The results show that, although public capital does not directly affect production in neighboring cantons, it does so indirectly by affecting production levels in its canton of origin, since evidence was found that production levels have a positive spatial correlation between cantons. In addition, evidence was found that the ability to propagate this effect does not depend solely on the size or economic relevance of the canton, but that there are additional characteristics that should be investigated.

The contribution of this study is relevant because, as far as the literature review has shown, it is the first approach in Ecuador to determine the spatial effects of the factors of production of a developing economy, and it also allows us to see which cantons propagate and receive these effects better. In addition, this study contributes to the academic discussion by showing evidence that the level of urban agglomeration is not the only factor that explains the capacity of a region to spill over its economic growth to its neighbors. Future analyses can delve deeper into the characteristics that make a canton

more likely to generate or receive these spatial effects.

The remainder of the paper is structured in the following manner. Section 2 sets the spatial production function framework and the model selection strategy. Section 3 gives a detailed description of the variables used in the model including its spatial autocorrelation analysis. Section 4 estimates the spatial models under different spatial dependence structures. Section 5 focus on the results of the selected model and presents direct and indirect output elasticities estimates. Section 6 gives some concluding remarks.

2 Spatial Production Function Model

According to Bom, Ligthart (2014), the base approach that has been used to analyze the effects of public capital consists of a Cobb-Douglas production function, which considers labor (L), public (G) and private (K) capital stocks in a function as factors of a region i that, when interrelated by a technological factor A, determine the aggregate production level Y_i:

$$Y_i = A_i L_i^{\beta_1} K_i^{\beta_2} G_i^{\beta_3}, \quad i = 1, \ldots, n \tag{1}$$

One of the main assumptions of this function is that the effects of public capital are directly related to the stock of public capital. For this case, the parameter of interest is β_3, which represents the partial elasticity of public capital production. This equation can be transformed to its log linear form by applying natural logarithm in the equation, which is convenient to perform an econometric analysis. For simplicity and in accordance with a general practice in the literature, it is assumed that the technological factor is equal to 1, in order to eliminate the direct influence of technology on the production function. This allows us to focus on the effect of capital and labor inputs. The equation is presented as follows:

$$\ln(Y_i) = \beta_1 \ln(L_i) + \beta_2 \ln(K_i) + \beta_3 \ln(G_i) \tag{2}$$

The analysis of the contribution of production factors on the productivity and income level of nations has been widely studied around the world. The neoclassical tradition has proposed the use of aggregate production functions, such as the Cobb-Douglas function, that explain the contribution of the components that contribute to the country's aggregate product (technology, capital and labor), through the analysis of their respective elasticities. According to Dall'erba, Llamosas-Rosas (2015), this function continues to be one of the most used ways to estimate production factors and technological progress.

In contemporary research, there is an increasing emphasis on understanding the spatial or interregional effects of public capital on production (Foster et al. 2023, Marrocu, Paci 2010). A spatial approach for studying economics affairs in Ecuador have been developed in recent years (Guevara-Rosero et al. 2019, Munoz, Pontarollo 2016, Szeles, Muñoz 2016). Their main focus have been on convergence and agglomeration phenomena.

Looking forward on this path, this research is based on the new economic geography perspective which proposes that economic entities, be they families or businesses, are spread out across diverse spatial locations, inherently separated by distances. This spatial dispersion instills the economy with a unique spatial structure that cannot be overlooked. Interactions among these entities tend to evolve, get delayed, or even get constrained by the physical distances between them. Similarly, there can be indirect or spatial economic ripple effects which might spread differently based on the degree of interconnectedness of these entities within a particular spatial framework.

2.1 Model selection

Based on LeSage, Pace (2009), LeSage, Fischer (2008), López-Bazo et al. (1999), Florax, Folmer (1992), Anselin, Rey (1991), Elhorst (2010), Munoz, Pontarollo (2016) summarises a strategy to model selection, it uses a (robust) Lagrange Multiplier (LM), likelihood ratio (LR) and a Wald test.

Following this suggested strategy, a spatial lag model was selected:

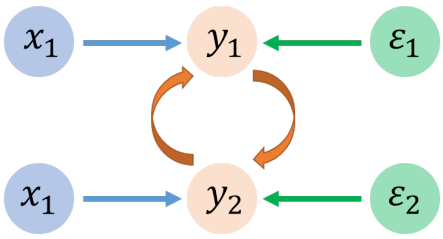


Figure 1: The spatial lag model for two regions. Straight lines represent non-spatial effects, curved lines are spatial effects

$$y = \rho W y + x\beta + \epsilon \tag{3}$$

where $y = \ln(Y)$ is a $n \times 1$ vector of observations of the dependent variable for n spatial units, ρ is the spatial autoregressive parameter which measures the intensity of the spatial interdependence, W is the $n \times n$ spatial weights matrix, β is a 3×1 coefficients vector of the covariates $\ln(L), \ln(K), \ln(G)$, and ϵ is the $n \times 1$ error term.

Figure 1 illustrates the spatial effects of two regions or spatial units in a spatial lag model. Golgher, Voss (2016) sets partial derivatives to study these effects (β_k coefficients represent the total effect of variable x_k):

$$S(W) = \begin{pmatrix} \frac{dy_1}{dx_{1k}} & \cdots & \frac{dy_1}{dx_{nk}} \\ \vdots & \ddots & \vdots \\ \frac{dy_n}{dx_{1k}} & \cdots & \frac{dy_n}{dx_{nk}} \end{pmatrix} = \beta_k (I - \rho W)^{-1} \tag{4}$$

where $S(W)_{11} = \frac{dy_1}{dx_{1k}}$ is the effect of x_k from region 1 over y of the same region and $S(W)_{n1} = \frac{dy_n}{dx_{1k}}$ is the effect of x_k from region 1 over y of region n. For a given covariate x_k, these let us define the average direct, total and indirect impacts:

$$\bar{M}_{\text{direct}} = n^{-1} \text{tr}(S(w) \tag{5}$$

$$\bar{M}_{\text{total}} = n^{-1} \iota_n^{-1} S(w) \iota_n \tag{6}$$

$$\bar{M}_{\text{indirect}} = \bar{M}_{\text{total}} - \bar{M}_{\text{direct}} \tag{7}$$

where ι_n is a $n \times 1$ vector of ones, $\bar{M}$ is the average effect.

Five spatial weights matrices W are applied with the chosen model. Contiguity matrices mark the elements of W with a dichotomous variable equal to 1 when the spatial units i and j are neighbors of each other and 0 otherwise. A *knn*-matrix based on a number k of nearest neighbors marks with 1 those regions that are within the k closest to each other. Specifically, we set three *knn*-matrices where $k = 5, 10$ and $k = 215$ (the total number of cantons minus 1). The inverse distance matrix W consists of dividing 1 for the distance weighting defined by the researcher. In this case, the greater the distance, the lower the weight assigned between regions.

3 Exploratory Spatial Data Analysis

3.1 The data

This study uses various public data sources to determine the dependent and independent variables for spatial regression analysis. Every data point in the dataset represents variables from 216 cantons within mainland Ecuador. Cantons without clear boundaries and those situated in the Galapagos Islands were not considered. Every canton is labeled using its unique code as per the National Institute of Statistics and Censuses (INEC) system.

Table 1: Data summary statistics. Per-capita values are shown in parentheses.

Statistic	NOGVA	Public	Private	WAP	Pop.
Min	5.20mn (858.4)	20,000 (0.88)	400 (0.015)	1,499	2,455
Q1	26.59mn (1,764.7)	668,148 (31.30)	2,820 (0.17)	8,848	13,085
Median	58.69mn (2,496.8)	1,632,026 (65.27)	31,320 (0.77)	18.760	28,080
Mean	421.50mn (3,155.4)	5,504,143 (93.49)	7,291,445 (19.52)	54,127	77,199
Q3	193.80mn (3,497.8)	4,679,626 (104.22)	565,097 (8.28)	39,856	60,519
Max	24.43mn (32,627.6)	183,876,079 (1,079.49)	539,377,575 (671.32)	1,943,861	2,644,891

Notes: "mn" ... million, "Pop." ... Population.

The geospatial data for the cantons was sourced from the Military Geographic Institute's (IGM) spatial database, which details Ecuador's territorial organization by cantons. This data was integrated into the primary database and employed to compute the spatial weight matrices for the model.

Table 1 shows summary statistics of the variables from year 2017 used in the study: non-oil gross value added (NOGVA), private investment (Private), public investment (Public), working age population (WAP) and population. Their per capita values are shown in parenthesis. We next provide a more in-depth explanation of the variables employed in the econometric modeling.

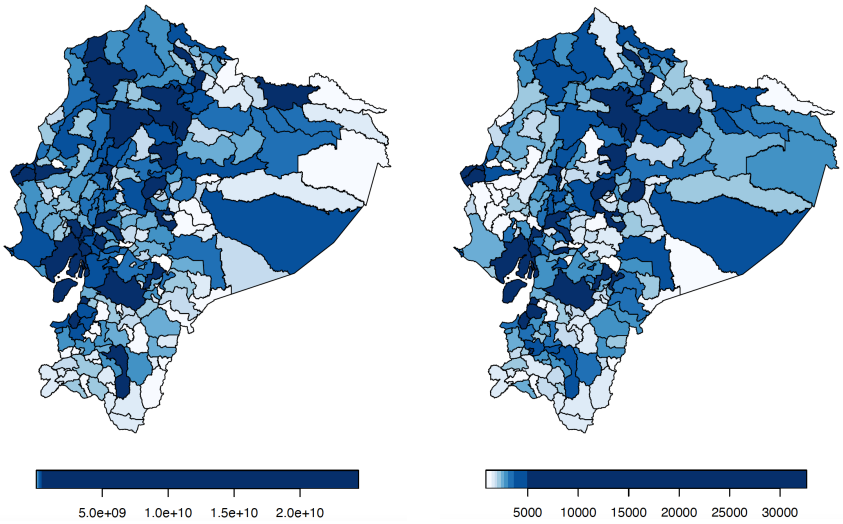
Production The non-oil gross value added variable is used, in per-capita terms for the year 2017 in US dollars ($NOGVApc$), obtained from the provisional regional accounts of the Central Bank of Ecuador as a proxy for production at the canton level. This variable was transformed into per-capita values with the population information from INEC. Figure 2a presents the spatial concentration of production in cantons: non-oil gross value added.

Public Capital Blades, Meyer-zu Schlochtern (1997) note that when it comes to specifying capital in productivity research, two main approaches are predominantly used:

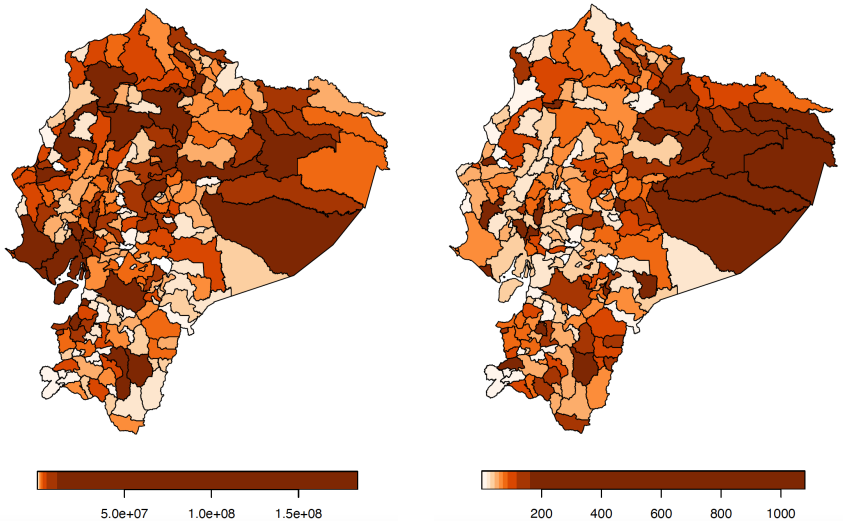
- When available in national accounts, the capital stock (CS) is used, signifying the capital assets' value within the economy. The gross capital stock (GCS) method values assets based on their acquisition time, ideal for calculating the total anticipated returns from assets over their lifespan. Yet, when gauging value-added changes for a single year, it is limited because it factors in projected income for the asset's entire useful life, both before and after the specified year. Conversely, the net capital stock (NCS) method omits projected income from years prior to the one under scrutiny but includes future anticipated earnings. The underlying rationale for these stock methods is the belief that capital services are aligned with its cost. Nevertheless, they overlook the fact that assets have diverse lifespans, meaning their production impact may vary within a particular year.
- Capital consumption (CC) over a specified timeframe serves as a proxy for discerning the capital contribution to the production function, especially for assets with diverse lifespans and years in operation. A notable downside is the inclusion of CC in production metrics like gross domestic product (GDP) or gross value added (GVA). Yet, these averages remain unaffected by capital consumption. This is because CC embodies the value that is subtracted to preserve the asset owner's wealth. Consequently, the author contends that annual fluctuations in GDP or GVA are not influenced by the CC.

Blades, Meyer-zu Schlochtern (1997) state that employing CC variables yields superior results compared to CS when analyzing Total Factor Production for the OECD, using 1999 data. This is attributed to the fact that the CC variable offers a more comprehensive insight into the growth of added value stemming from the capital factor's contribution.

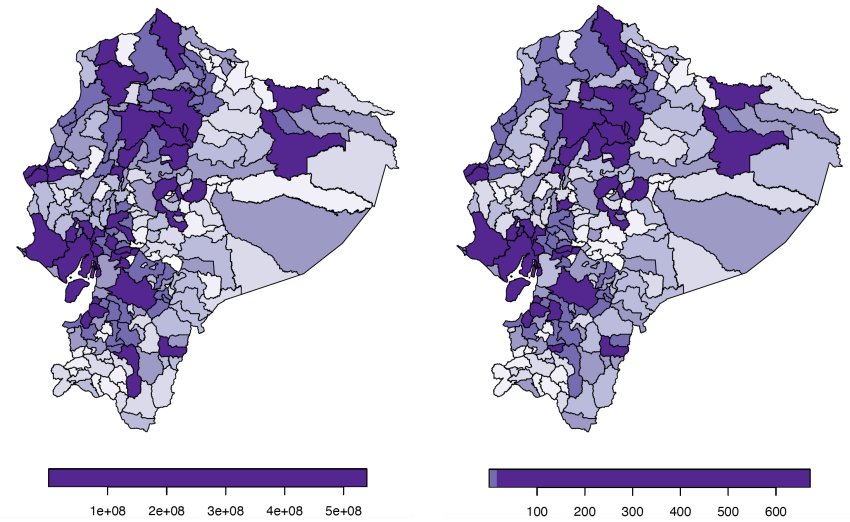
For Ecuador, cantonal-level data for CS or CC variables, like the gross fixed capital formation (GFCF) related to public capital, is absent. Consequently, in alignment with



(a) Non-oil gross value added (left panel) and per-capita values (right panel)



(b) Public capital (left panel) and per-capita values (right panel)



(c) Private capital (left panel) and per-capita values (right panel)

Figure 2: Spatial distributions (Choropleth maps) of main variables

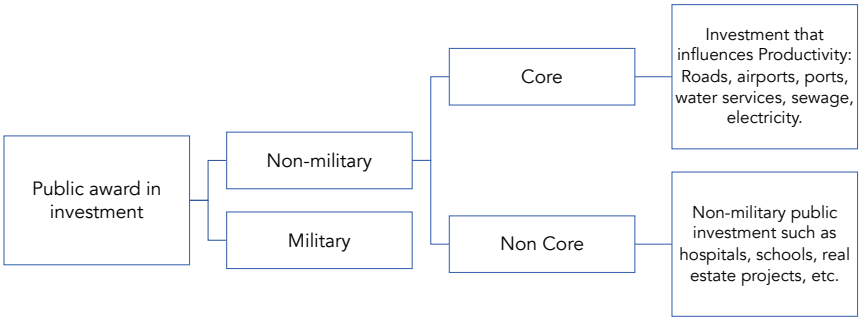


Figure 3: Public investment classification

employing a CC-based approach as a proxy for public capital, data from the National Public Procurement Service (SERCOP in Spanish) from 2017 is used.

The data entries in this source are recorded at the process level of contracting. However, they do not include variables specifying the canton where the work takes place. Yet, each data point has an identifier for the contracting entity responsible for the award, and this identifier includes the Entity's RUC (Unique Registry of Taxpayers).

In an effort to identify the location of various projects, a variable was created using the RUC of the awarding public entities. These recorded work data points were then matched with the Fiscal Administration (SRI in Ecuador) RUC database, which provides information about the canton where each entity is based. This merge resulted in an intermediate dataset detailing awarded contracts along with the respective canton of each entity. However, this dataset only indicates the location of the contracting entity and not necessarily the exact canton where the work occurs. This distinction is particularly important for contracting entities that invest in multiple cantons beyond their primary location. This is especially true for entities like the Decentralized Autonomous Governments at both national and provincial levels and regional electrical companies. To illustrate, the Decentralized Autonomous Government of Azuay, headquartered in the provincial capital of Cuenca, oversees projects not just in Cuenca but in other cantons within that province. Given this complexity, a meticulous case-by-case review was essential to accurately assign the correct canton to each contracting process. This involved in-depth analysis of individual contracting processes to pinpoint the specific canton for each investment. Nonetheless, for Decentralized Autonomous Governments at the cantonal and parish levels, and their public corporations, such scrutiny was not required. Their projects are typically located in the same canton as the entity's main office.

Furthermore, in line with existing literature, these projects were categorized as either non-military or military and also delineated between core and non-core (Figure 3)

In the final step, the data pertaining to the amounts awarded by canton were incorporated, with a focus on exclusively including those related to core public capital projects. This process resulted in the creation of a variable containing the award amounts for core public works, organized by canton and expressed in US dollars. It is transformed into per capita terms using the INEC population projection for the year 2017, which was prepared with data from the 2010 census, calling this variable *PubCpc* which canton concentration is shown in Figure 2b.

Private Capital To represent private capital, data on corporate capital expenditures from the Superintendency of Companies of Ecuador were utilized. This data, available at the canton level, was then aggregated per canton and converted into per capita terms (PrivCpc). Canton concentration is shown in Figure 2c.

Labor For the effects of labor's role within the Cobb-Douglas function, and to align it in US dollar terms like the other variables, the method proposed by Han et al. (2016) was adopted. This method equates labor to the Economic Working Age Population (WAP). To achieve this, population projections from the 2010 census were utilized. These

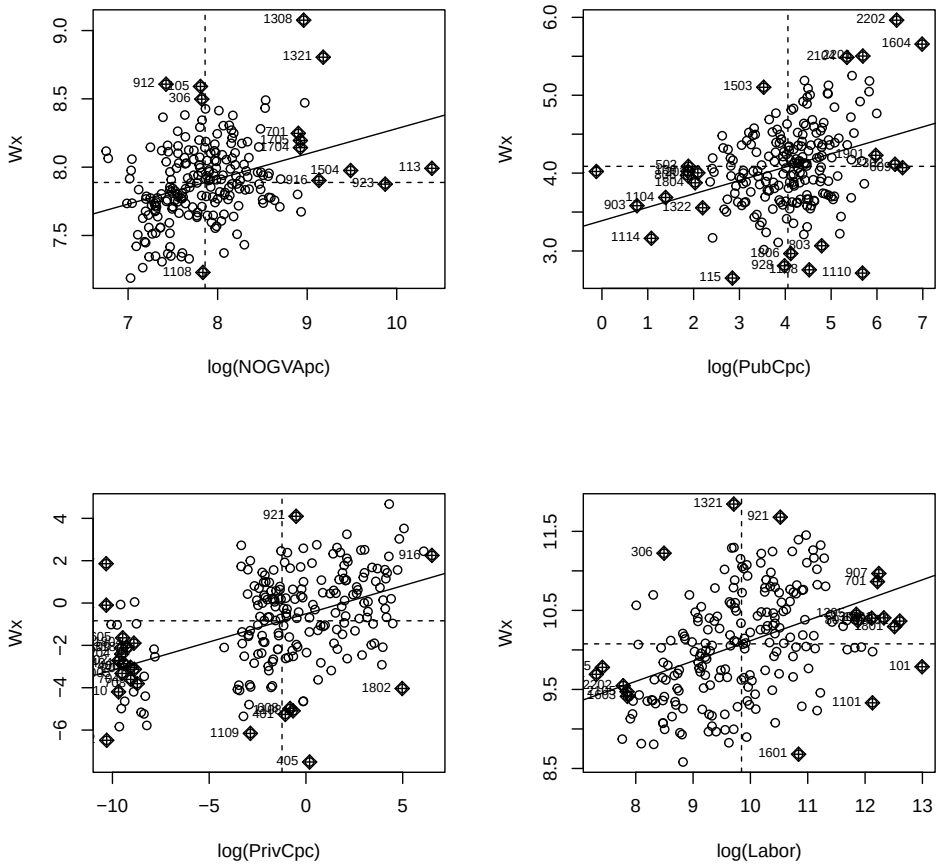


Figure 4: Moran plot for the logarithms of non-oil gross value added (NOGVApc), public capital per capita (PubCpc), private capital per capita (PrivCpc) and labor.

projections are sorted by canton and age. Subsequently, data from each canton regarding the population aged 15 and above was aggregated, aligning with the WAP definition.

3.2 Spatial Autocorrelation

Moran's I test is utilized in order to test for spatial dependency. The assessment is based on a hypothesis that is a random spatial distribution of the observations. If the null hypothesis is rejected, it suggests that there is a discernible spatial pattern or structure embedded within the data.

Figure 4 shows positive Moran's I for the logarithms of non-oil gross value added (NOGVApc), public capital per capita (PubCpc), private capital per capita (PrivCpc) and labor. They are all significant at 5% which is confirmed in Table 2. They suggest underlying spatial dependence in all variables. The Moran plot's first and third quadrants (high-high, HH, and low-low, LL) display cantons that are neighbored by other cantons with similar values, whether consistently high (in the case of HH) or consistently low (for LL). The second and fourth quadrants of the Moran plot, namely low-high (LH) and high-low (HL), exhibit cantons where a low (or high) value of the variable is neighbored by cantons with high (or low) values of the same variable. Cantons are present in all quadrants of Figure 4. Quadrants I and III have over 60% of cantons which explains positive slopes.

Table 2 presents the Moran's I statistic (MI), its expected value (E[MI]), variance (V[MI]), z-value and p-value under different approaches for variance computation: Randomization, Normal and Monte Carlo. Z-value let us compare across these setups. In the

Table 2: Moran's I test for the logarithms of non-oil gross value added (NOGVApc), public capital per capita (PubCpc), private capital per capita (PrivCpc) and labor.

	log(NOGVApc)			log(PubCpc)		
	Rnd.	**Normal**	**MC**	**Rnd.**	**Normal**	**MC**
MI[a]	0.2087	0.2087	0.2087	0.1662	0.1662	0.1662
E[MI][b]	-0.0047	-0.0047	-0.0042	-0.0047	-0.0047	-0.0041
V[MI][c]	0.0018	0.0018	0.0020	0.0018	0.0018	0.0020
z-value	4.9898	4.9645	4.7512	3.9911	3.9772	3.8430
p-value	0.0000	0.0000	0.0010	0.0000	0.0000	0.0010

	log(PrivCpc)			log(Labor)		
	Rnd.	**Normal**	**MC**	**Rnd.**	**Normal**	**MC**
MI[a]	0.2850	0.2850	0.2850	0.2622	0.2622	0.2622
E[MI][b]	-0.0047	-0.0047	-0.0041	-0.0047	-0.0047	-0.0049
V[MI][c]	0.0018	0.0018	0.0017	0.0018	0.0018	0.0018
z-value	6.7409	6.7415	6.9179	6.2241	6.2100	6.3173
p-value	0.0000	0.0000	0.0010	0.0000	0.0000	0.0010

Notes: [a]Moran's I Statistic; [b]Expected Moran's I; [c]Moran's I variance; "Rnd.": Randomization; "MC": Monte Carlo.

case of non-Oil gross value added (NOGVApc) and public capital per capita (PubCpc), Moran's I is greatest under randomization (4.9898 and 3.9911 respectively). For private capital per capita (PrivCpc) and labor, Moran's I is greatest in Monte Carlo (6.9179 and 6.3173 respectively). All results implies that there is evidence of robust positive spatial autocorrelation at 5% significance level in all cases.

4 Spatial Model choice

As mentioned in Section 2.1, Lagrange multiplier (LM), likelihood ratio (LR) and a Wald test are used to select the spatial lag model. Table 3 presents both the LM test statistics and the robust LM test statistics, specifically for a spatial lag in the dependent variable and for a spatial error term. Accompanying these statistics are the respective p-values. Versions that are not robust show significant p-values but robust counterparts do not.

Testing with different spatial matrices allows researchers to study spatial sensitivity. Each type of matrix captures a distinct notion of spatial interaction—for example, contiguity matrices focus on neighboring units, while distance-based matrices emphasize proximity, and k-nearest neighbor matrices ensure each unit is connected to a fixed number of others. By examining results across different spatial weight specifications, analysts can assess whether spatial dependence remains consistent under varying definitions of spatial proximity (Anselin, Rey 1991). In our case, spatial sensitivity to changes in spatial weights is examined through two approaches: Lagrange Multiplier Tests and the estimation of the SLM coefficients.

Table 3 outlines different weights matrices. Contiguity weight matrix is a standard base approach, *knn* distance matrices (*knn* 5, 10, 215) let us examine the robustness of the estimation as more neighbors are included. Inverse distance let us check the behavior of model estimation *inverting* the weights as distance is greater. It is worth noting that, for regularity conditions, all weights are row-normalized. These results exhibit a clear pattern: spatial weights matrices that emphasize closer relationships yield significant p-values, while those representing broader or more distant spatial interactions tend to produce insignificant p-values as the spatial range increases.

Traditional LM tests, considering all contiguity and *knn* (up to k=10) spatial weights matrices, show consistent results in the sense that they reject the hypothesis of no spatially lagged-dependent variable at a 5% significance level. However, robust LM tests the hypothesis of no spatially autocorrelated error is not rejected for any spatial weights matrix. As illustrated in Figure 2 of Putra et al. (2020), when the Robust LM test is not significant, no clear decision can be made. In this case, the LR and Wald tests can assist in determining the appropriate model.

Table 3: Lagrange multiplier tests for a spatially lagged-dependent variable and spatial error correlation.

	LMlag		RLMlag		LMerr		RLMerr	
	Statistic	**p-value**	**Statistic**	**p-value**	**Statistic**	**p-value**	**Statistic**	**p-value**
Contiguity	13.977	0.000	0.635	0.426	14.084	0.000	0.742	0.389
knn 5	12.357	0.000	0.036	0.850	14.788	0.000	2.467	0.116
knn 10	4.346	0.037	0.235	0.628	6.887	0.009	2.776	0.096
knn 215	0.502	0.478	0.000	1.000	0.502	0.478	0.000	1.000
Inverse distance	1.385	0.239	0.236	0.627	2.320	0.128	1.170	0.279

Table 4: p-values from likelihood ratio (LR) and a Wald test. Columns SDM, SLM and SEM show LR of row and column comparison

p-values	**SDM**	**SLM**	**SEM**	**Wald**	**LR**
SAC[a]	0.8175	0.6794	0.9041	0.7246	0.0141
SDM[b]		0.7101	0.8335	0.0001	0.0003
SLM[c]			-	0.0002	0.0004
SEM[d]				0.0001	0.0003

Notes: [a]Spatial autoregressive model; [b]Spatial Durbin model; [c]Spatial lag model; [d]Spatial error model.

Table 4 present p-values from LR and Wald tests in the last two columns. The null hypothesis in these cases is the absence of spatial dependence, the hypothesis is rejected in almost all models except in SAC for Wald test. The first three columns in Table 4 show LR p-values of row-column model specifications: spatial autoregressive model (SAC), spatial Durbin model (SDM), spatial lag model (SLM) and spatial error model (SEM). For example, 0.7101 is the LR test p-value of comparing SDM and SLM. This table shows there is no difference, it reduces our model specification to SLM and SEM based on the parsimony principle.

Although the SEM model considers spatial dependence in the disturbance process, it does not offer insights into spillovers (Elhorst, Vega 2013). As our goal is to investigate the impact of public capital spillovers on gross value added, and the available evidence supports the use of spatial lag model (SLM), it is the preferred method over spatial error model (SEM).

5 Results

5.1 Estimation and Impacts

We examine if the production level of a canton can impact the corresponding variable in its adjacent cantons. Estimation results of SLM (spatial lag model) are presented in Table 5. There are 6 models depending on the spatial weights specification: (0) ordinary least squares-OLS (1) geographical contiguity, (2) k-nearest neighbors with $k = 5$, (3) k-nearest neighbors with $k = 10$, (4) k-nearest neighbors with $k = 215$, and (5) inverse distance.

In linear regressions, including spatial linear regressions, conclusions about the significance of the coefficients can be misleading in the presence of multicollinearity (Corrado, Fingleton 2012). Based on Morales-Oñate, Morales-Oñate (2023), a multicollinearity test was performed in OLS model finding that the multicollinearity hypothesis is rejected at 5% significance for all variables.

The findings indicate a positive spatial correlation among the production levels (GVA) of various cantons in Ecuador. This is evident in the significant ρ value observed for the contiguity and neighborhood matrices up to closest 10. However, this is not the case for other spatial weights specifications.

Based on Kubara, Kopczewska (2024), it was determined that setting $k = 4$ optimizes the Akaike information criterion (AIC), yielding a value of $AIC = 291.5433$. Furthermore, the study suggests that fine-tuning W by adjusting a few spatial units (such as changing knn from 5 to 4) result in negligible gains, consistent with our findings. Among

Table 5: Estimation results in spatial lag model.

	OLS (Model (0))		**Contiguity** (Model (1))		**knn 5** (Model (2))	
	Estimate	**p-value**	**Estimate**	**p-value**	**Estimate**	**p-value**
(Intercept)	6.356	0.000	3.959	0.000	3.883	0.000
log(PubCpc)	0.099	0.002	0.094	0.002	0.093	0.002
log(PrivCpc)	0.036	0.000	0.031	0.001	0.036	0.000
log(Labor)	0.117	0.000	0.109	0.000	0.106	0.001
ρ			0.315	0.000	0.330	0.001
Log-likelihood	-146.7143		-140.3441		-140.703	
AIC	301.4286		292.6883		293.4061	

	knn 10 (Model (3))		**knn 215** (Model (4))		**Inverse distance** (Model (5))	
	Estimate	**p-value**	**Estimate**	**p-value**	**Estimate**	**p-value**
(Intercept)	4.344	0.000	14.219	0.199	2.198	0.304
log(PubCpc)	0.098	0.002	0.098	0.002	0.097	0.002
log(PrivCpc)	0.035	0.000	0.036	0.000	0.036	0.000
log(Labor)	0.112	0.000	0.116	0.000	0.113	0.000
ρ	0.262	0.047	-0.998	0.246	0.211	0.211
Log-likelihood	-144.7379		-146.0412		-145.9331	
AIC	301.4759		304.0823		303.8663	

all *knn* distance matrices in Table 5, *knn* 5 emerges as the optimal. Following the AIC criteria, the contiguity matrix has the lowest AIC overall. In accordance with the AIC criteria, the contiguity matrix exhibits the lowest AIC overall.

If we were to base our selection solely on the specification of W, the contiguity matrix would be the preferred option. However, our objective is to present all possible scenarios whenever feasible. The analysis compares different spatial weight matrices to test the robustness of SLM and assess how various spatial structures affect estimated spatial effects. Although the contiguity matrix was used for the main analysis, k-nearest neighbors (k=5,10,215) and inverse distance matrices helped validate the results. The consistency of spatial lag coefficients and p-values across different matrices confirmed the stability of the findings (LeSage, Pace 2009). This comparison enhances model credibility and contributes to refining spatial weight matrix selection in future research.

When working with a geographically incomplete dataset, the concept of contiguity might not be suitable. In our case, four insular cantons and two cantons from Guayas (General Antonio Elizalde) and Manabi (Junin) were removed due to lack of information. In Continental Ecuador, we work with 99.08% of cantons. Therefore, we can reasonably assume our dataset as complete.

The estimates of Model (3) are slightly higher than the coefficients in Models (1) and (2), coefficients of Model (0) are the highest. Estimated coefficients of public, private and labor variables are significant at 5% in almost all cases. ρ in Model (1) and Model (2) are significant, large and similar, it decreases and loses significance in the rest of the models. It is not appropriate to compare the coefficient estimates of spatial models to OLS, as the coefficient estimates in spatial models exclusively capture the direct marginal effects. We obtain mean direct effects, mean indirect effects, and total effects for comparison purposes. It is worth noting that ρ decreases as distance grow. This is consistent with ρ which is not significant in the inverse distance spatial weights specification.

Upon identifying evidence of an indirect spatial effect between the production levels of the cantons, our focus shifted to quantifying the influence exerted by the production factors via this transmission mechanism. Table 6 showcases the direct and indirect output elasticity calculations, which are derived from the coefficient estimates found in Table 5.

Utilizing the S matrix in equation (4), we discovered significant evidence supporting these indirect effects. Specifically, the average indirect effect of public capital, when evaluated with contiguity, stands at 13.69% with significance level at 5%. In comparison, private capital manifests a slightly more pronounced impact at 4.12%, and labor displays the most substantial indirect effect, measuring 4.77%. Similar results are obtained for distance up to five neighbors. However, significance of indirect impacts is lost in the rest

Table 6: Direct and indirect output elasticity estimates.

	Contiguity (Model (1))		**knn 5** (Model (2))		**knn 10** (Model (3))		**knn 215** (Model (4))		**Inv. dist.** (Model (5))	
	Est.	**p-val.**	**Est.**	**p-val.**	**Est.**	**p-val.**	**Est.**	**p-val.**	**Est.**	**p-val.**
log(PubCpc)										
Total	0.1369	0.0033	0.1388	0.0046	0.1330	0.0071	0.0027	0.6015	0.2091	0.3832
Direct	0.0958	0.0016	0.0949	0.0023	0.0989	0.0017	0.0984	0.0010	0.0979	0.0018
Indirect	0.0412	0.0478	0.0439	0.0484	0.0342	0.1674	-0.0958	0.8371	0.1111	0.5434
log(PrivCpc)										
Total	0.0453	0.0015	0.0532	0.0004	0.0478	0.0014	0.0010	0.5988	0.0764	0.3984
Direct	0.0317	0.0007	0.0364	0.0002	0.0356	0.0003	0.0355	0.0003	0.0358	0.0002
Indirect	0.0136	0.0362	0.0168	0.0218	0.0123	0.1331	-0.0346	0.8307	0.0406	0.5588
log(Labor)										
Total	0.1586	0.0010	0.1588	0.0016	0.1519	0.0023	0.0031	0.5951	0.2421	0.3648
Direct	0.1109	0.0003	0.1086	0.0008	0.1129	0.0006	0.1164	0.0005	0.1134	0.0005
Indirect	0.0477	0.0353	0.0503	0.0306	0.0390	0.1331	-0.1133	0.8351	0.1287	0.5337

Notes: "Est.": Estimate; "p-val.": p-value; "Inv. dist.": Inverse distance.

of spatial weights matrix specifications.

Taking into account the total effect of public capital on economic performance in Model (1), which is 0.045, and breaking it down into its components (direct: 0.0317 and indirect: 0.0136), we find that the spatial (indirect) component accounts for 30% of the overall impact. Meanwhile, the direct effect contributes the remaining 70%. To determine the feedback effects of each factor input, we subtract the coefficient estimates from the direct output elasticity estimates. For example, in the case of public capital, the feedback effect is $0.0958 - 0.094 = 0.002$. This means that each canton exerts a feedback effect of 0.002 on its neighboring cantons, which in turn influences their neighbors, creating a ripple effect throughout the network. For private capital and labor, the feedback effect is 0.001 and 0.002 respectively.

The findings suggest that public capital, along with other production factors, produces spatial impacts among adjacent cantons. This chain of influence stems from how these factors affect external production levels, which subsequently shape the production levels of neighboring spatial entities.

5.2 Marginal effects by cantons

In subsection 5.1, direct, indirect and total effects were computed and analyzed. These measures give us valuable average information. However, the average indirect effects fail to convey the spillover impacts of individual canton on one another. Fixing the estimated public capital coefficient $\beta = 0.0938$ in equation (4), generates a $S(W)$ matrix of size $n \times n$ whose elements let us capture these individual canton spillovers. A unit increment of public capital in canton i has an individual direct effect on production of the same canton i (diagonal of $S(W)$). Also, a unit increment of public capital in canton i has an individual indirect effect on production of canton j (off-diagonal of $S(W)$). We are interested in the row and column sums of the off-diagonal elements of $S(W)$, $i \neq j$ which are called row effects and column effects respectively.

Leveraging the spatial contiguity weights matrix and fixing the estimated public capital coefficient in equation (4), we delve into the spatial impacts of public capital on individual cantons. We dissect both the row and column effects to determine which spatial units exert the most influence over their adjacent counterparts (column effects: total spillover effects of a specific canton onto the production of other cantons) and identify which units are more reliant on their neighboring regions (row effects: when all other cantons increase public capital input by one unit, row effects are spillover effects from other cantons to a specific canton). Figure 6, Table 7 and Table 8 show these effects.

The findings highlight that the Cañar canton is the preeminent canton in the country that positively impacts its neighbors through public investment. It is crucial to note that this canton has two unique interior neighbors, which exclusively share a border with Cañar (see Figure 5). Among Suscal, Cañar and El Tambo, Cañar has 28% of gross

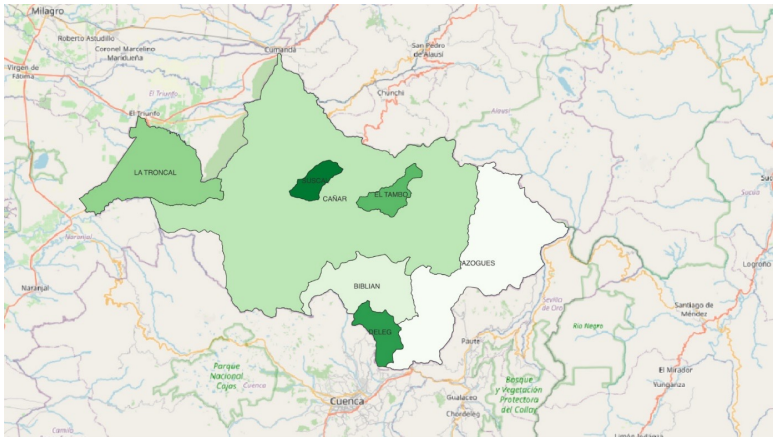


Figure 5: Cañar

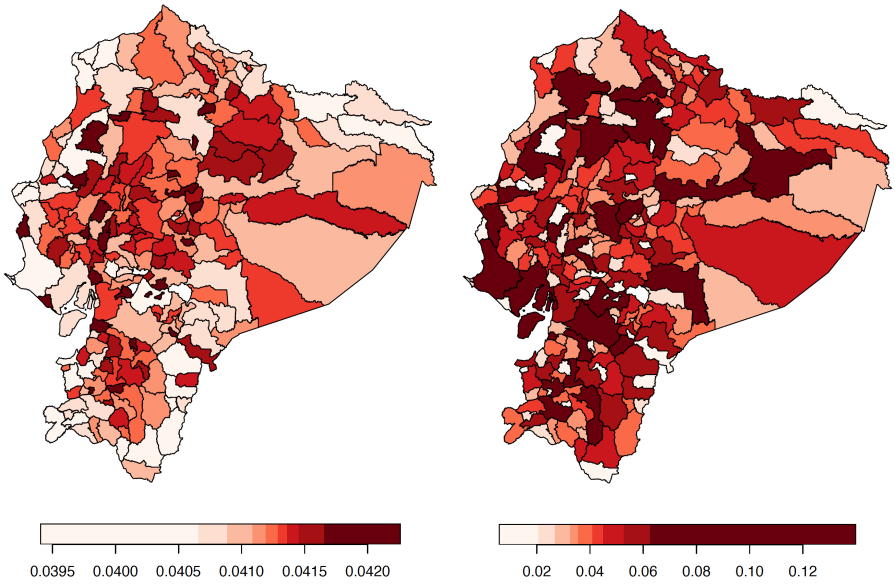


Figure 6: Public's capital row (left) and column (right) effects on per capita non-oil gross value added

value added (GVA), 23% of public capital and 79% of the population. This suggests that it preeminent column effect is influenced by a population effect. Ecuador's major cities – Quito, Cuenca, and Guayaquil – belong to the primary top 10 cantons where public investment significantly affects surrounding areas. Nonetheless, Table 7 also presents ranked population size and GVA, which seem not to be decisive factors in determining the observed impact of public investment since their log-scale Pearson correlation with column effects are 0.37 and 0.40, respectively. Spatial structure play a significant role in this regard since the contiguity weight matrix indicates that 60% of Ecuador's cantons have five to 12 neighbors.

Column effects in Table 7 can be interpreted as follows. On average, an increase of one percentage point in public capital in the Santa Elena canton increases the economic performance (measured in terms in GVA) of its surrounding cantons by 0.1036%.

On the other hand, Table 8 (row effects) show the cantons that benefit most from the public investment of their neighbors, which are Tambo and Suscal. They are completely surrounded by the Cañar canton, which generates the greatest column effect. Row effects in Table 8 can be interpreted as follows. In the case of the Rumiñahui canton, on average, an increase of one percentage point in public capital in its surrounding cantons increases its economic performance by 0.0419%.

Table 7: Public's capital column effects on per-capita non-oil gross value added.

Column effect		GVA		Population		Canton		
Rank	Value	Rank	Value	Rank	Value	Code	Canton	Province
1	0.1398	55	192,390,383	49	66,996	303	CAÑAR	CAÑAR
2	0.1053	144	36,163,641	122	24,017	1109	PALTAS	LOJA
3	0.1036	27	417,373,082	17	176,373	2401	SANTA ELENA	SANTA ELENA
4	0.1020	1	24,426,597,900	2	2,644,145	1701	D.M. QUITO	PICHINCHA
5	0.0973	35	314,327,442	21	131,877	1303	CHONE	MANABI
6	0.0952	3	4,392,835,893	3	603,269	101	CUENCA	AZUAY
7	0.0921	2	20,554,798,446	1	2,644,891	901	GUAYAQUIL	GUAYAS
8	0.0874	16	905,261,666	18	171,038	1201	BABAHOYO	LOS RIOS
9	0.0838	42	255,159,287	45	74,158	1501	TENA	NAPO
10	0.0788	20	655,491,210	20	140,670	804	QUININDE	ESMERALDAS

Table 8: Public's capital row effects on per-capita non-oil gross value added.

Row effect		GVA		Population		Canton		
Rank	Value	Rank	Value	Rank	Value	Code	Canton	Province
1	0.0423	113	56,500,676	168	11,673	305	EL TAMBO	CAÑAR
2	0.0423	186	15,269,624	200	6,128	307	SUSCAL	CAÑAR
3	0.0421	78	108,585,168	61	54,308	921	PLAYAS	GUAYAS
4	0.0420	130	44,499,752	116	24,615	1305	FLAVIO ALFARO	MANABI
5	0.0420	153	30,696,591	164	12,982	605	CHUNCHI	CHIMBO-RAZO
6	0.0419	18	803,979,272	25	107,043	1705	RUMIÑAHUI	PICHINCHA
7	0.0419	107	59,324,110	115	24,777	903	BALAO	GUAYAS
8	0.0419	141	41,454,460	126	23,689	1319	PUERTO LOPEZ	MANABI
9	0.0418	73	121,913,902	68	50,241	2302	LA CONCORDIA	S.T. DE LOS TSACHILAS
10	0.0418	9	1,484,310,229	7	293,005	907	DURAN	GUAYAS

It is interesting that in the lists of the main cantons there are several satellite cities, such as Rumiñahui, which borders the Metropolitan District of Quito and Durán with Guayaquil.

The large cities in Ecuador, such as Quito and Guayaquil, concentrate a large part of the country's economic activity, and the surrounding cantons are usually home to workers and companies that interact with these economic centers and benefit from their economic dynamism. Therefore, an increase in the economic activity of these cities linked to public capital investments can have a significant influence on the surrounding cantons.

However, it can be argued that the political-administrative power of these cities can also influence the economic performance of the surrounding regions. Therefore, it should be considered that in Ecuador each of these regions has legal autonomy over its competencies, therefore, when analyzing regions with the same level of hierarchy (cantons) there can be no inference in the political decision making of larger cantons. Additionally, although Quito and Guayaquil are considered metropolitan districts in Ecuador, they do not contain other municipalities or cantons within them, as is the case with most metropolitan districts worldwide.

There are several mechanisms for the transmission of spillover effects of public capital between cantons. As explained by Berndt, Hansson (1992), public capital can reduce firms' production costs, improving their output and performance. This in turn motivates firms to demand goods, services and labor from neighboring cantons, thereby increasing household production and consumption.

This increase in productivity can also encourage the formation of industrial clusters. These clusters can expand to neighboring cantons, as has been the case of Rumiñahui, which is a satellite canton of Quito, or Durán, which is located near Guayaquil.

In addition to these causes, investment in connectivity infrastructure can improve access to services in neighboring cantons, as well as boost trade and labor mobility, which impacts production in neighboring cantons.

6 Conclusions

Similar to the literature found on developed countries, this research has found evidence of public capital spillover effects in Ecuador. To the best of our knowledge, there are no clear studies differentiating between public and private (capital) spillovers on developing countries including Ecuador. Our work can give a guidance to follow a similar path in information gathering about public capital, exploring spatial structures and elasticity analysis to be explored in future research.

The findings indicate that in Ecuador, production factors, especially public capital, establish spatial relationships among the cantons. This primarily transmission mechanism is through the production levels within the cantons themselves. The SLM model evaluated with a contiguity matrix shows that the spatial effects of public capital (0.012) can explain 30% of the total effect that this factor has on the economic performance of the cantons. In contrast, the non-spatial or direct influence (0.032) represents 70%. Given its significance in the total impact, the spatial structure in the model is essential, suggesting that it is not feasible to assume independence among the cantons under study.

Although the SLM model indicates that the most populous cities in Ecuador have the most substantial direct and indirect effects on their neighboring cantons, there are also smaller cities, both in terms of population and economic significance, that play a role in this dynamic.

In addition, this study contributes to the academic discussion by showing evidence that the level of urban agglomeration is not the only factor that explains the capacity of a region to spill over its economic growth to its neighbors. By taking a more detailed sample of regions, a more specific analysis of the possible economic and spatial dynamics that arise between them should be carried out.

The findings have important implications for shaping public policies, especially those directed at promoting regional growth and development. These implications arise from the ability to direct investments preferentially towards cantons that demonstrate a more significant regional ripple effect. Nevertheless, any policy formulation should also consider the temporal dynamics of these effects to ensure enduring and equitable growth across regions.

Future research could delve into the longitudinal variation of these effects, probing how they evolve over extended periods. Additionally, a more granular examination could be undertaken to discern the specific attributes that lead certain cantons to exert a more pronounced contagion influence, as well as to identify which cantons derive the most significant benefits from these ripple effects.

References

Álvarez IC, Barbero J, Zofío JL (2016) A spatial autoregressive panel model to analyze road network spillovers on production. *Transportation Research Part A: Policy and Practice* 93: 83–92. CrossRef.

Anselin L, Rey S (1991) Properties of tests for spatial dependence in linear regression models. *Geographical analysis* 23[2]: 112–131. CrossRef.

Aschauer DA (1989) Is public expenditure productive? *Journal of monetary economics* 23[2]: 177–200. CrossRef.

Berndt ER, Hansson B (1992) Measuring the contribution of public infrastructure capital in Sweden. *The Scandinavian Journal of Economics* 94: 151–168. CrossRef.

Blades D, Meyer-zu Schlochtern J (1997) How should capital be represented in studies of total factor productivity. Canberra group on capital stock statistics, OECD, https://www.oecd.org/sdd/na/canberragrouponcapitalstockstatistics-march1997.htm

Bom PR, Ligthart JE (2014) What have we learned from three decades of research on the productivity of public capital? *Journal of economic surveys* 28[5]: 889–916. CrossRef.

Briones Bendoza XF, Molero Oliva LE, Zamora OX (2018) The production function Cobb-Douglas in Ecuador. *Tendencias* 19[2]: 45–73. CrossRef.

Corrado L, Fingleton B (2012) Where is the economics in spatial econometrics? *Journal of Regional Science* 52[2]: 210–239

Dall'erba S, Llamosas-Rosas I (2015) The impact of private, public and human capital on the US states economies: Theory, extensions and evidence. *Handbook of research methods and applications in economic geography*: 436–467. CrossRef.

Elhorst JP (2010) Applied spatial econometrics: Raising the bar. *Spatial economic analysis* 5[1]: 9–28. CrossRef.

Elhorst P, Vega SH (2013) On spatial econometric models, spillover effects, and W. 53rd congress of the European Regional Science Association, Palermo, Italy, http://hdl.handle.net/10419/123888

Fingleton B (2001) Equilibrium and economic growth: Spatial econometric models and simulations. *Journal of regional Science* 41[1]: 117–147. CrossRef.

Florax R, Folmer H (1992) Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional science and urban economics* 22[3]: 405–432. CrossRef.

Foster V, Gorgulu N, Straub S, Vagliasindi M (2023) The impact of infrastructure on development outcomes: A qualitative review of four decades of literature. Policy Research Working Papers, 10343, World Bank, http://hdl.handle.net/10986/39515

Golgher AB, Voss PR (2016) How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spatial Demography* 4: 175–205. CrossRef.

Guevara C (2016) Growth agglomeration effects in spatially interdependent Latin American regions. *Journal of Applied Economic Sciences* 11[45]: 1350–1355

Guevara-Rosero GC, Riou S, Autant-Bernard C (2019) Agglomeration externalities in Ecuador: Do urbanization and tertiarization matter? *Regional Studies* 53[5]: 706–719. CrossRef.

Han J, Ryu D, Sickles R (2016) How to measure spillover effects of public capital stock: a spatial autoregressive stochastic frontier model. In: Baltagi BH, Lesage JP, Pace RK (eds), *Spatial econometrics: Qualitative and limited dependent variables*. Emerald, 259–294. CrossRef.

Jia K, Qiao W, Chai Y, Feng T, Wang Y, Ge D (2020) Spatial distribution characteristics of rural settlements under diversified rural production functions: A case of Taizhou, China. *Habitat International* 102: 102201. CrossRef.

Kubara M, Kopczewska K (2024) Akaike information criterion in choosing the optimal k-nearest neighbours of the spatial weight matrix. *Spatial Economic Analysis* 19[1]: 73–91. CrossRef.

LeSage J, Pace RK (2009) *Introduction to spatial econometrics*. Chapman and Hall/CRC. CrossRef.

LeSage JP, Fischer MM (2008) Spatial growth regressions: model specification, estimation and interpretation. *Spatial Economic Analysis* 3[3]: 275–304. CrossRef.

López-Bazo E, Vayá E, Mora AJ, Suriñach J (1999) Regional economic dynamics and convergence in the European Union. *The Annals of Regional Science* 33[3]: 343–370. CrossRef.

Marrocu E, Paci R (2010) The effects of public capital on the productivity of the Italian regions. *Applied Economics* 42[8]: 989–1002. CrossRef.

Mera K (1973) Regional production functions and social overhead capital: An analysis of the Japanese case. *Regional and Urban Economics* 3[2]: 157–185. CrossRef.

Morales-Oñate V, Morales-Oñate B (2023) MTest: A bootstrap test for multicollinearity. *Revista Politécnica* 51[2]: 53–62. CrossRef.

Moreno Loza JL (2017) Incidencia del aumento sostenido del gasto público sobre el PIB en Ecuador: multiplicador fiscal para el período 2000–2015. B.S. thesis, PUCE

Munoz RM, Pontarollo N (2016) Cantonal convergence in Ecuador: A spatial econometric perspective. *Journal of Applied Economic Sciences* 11[1]: 39

Putra AS, Tong G, Pribadi DO (2020) Spatial analysis of socio-economic driving factors of food expenditure variation between provinces in Indonesia. *Sustainability* 12[4]: 1638. CrossRef.

Ram R (1996) Productivity of public and private investment in developing countries: A broad international perspective. *World Development* 24[8]: 1373–1378. CrossRef.

Szeles MR, Muñoz RM (2016) Analyzing the regional economic convergence in Ecuador. Insights from parametric and nonparametric models. *Romanian Journal of Economic Forecasting* 19[2]: 43–65

REGION
The Journal of ERSA
Powered by WU

Volume 12, Number 1, 2025, 19–45
DOI: 10.18335/region.v12i1.563

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

Spatial Inequality

Sergio J. Rey[1]

[1] San Diego State University, San Diego, USA

Received: August 28, 2024/Accepted: February 7, 2025

Abstract. This paper explores the concepts and computational methods used to measure spatial inequality, emphasizing a reproducible approach that social scientists can apply to their research. The analysis focuses on geographic income disparities at the sub-national level, using Mexico as a case study. By examining various a-spatial and spatially explicit approaches, the paper highlights the complexities of measuring inequality across places and over time. The discussion includes a review of traditional inequality measures and introduces spatial decomposition methods that account for the geographical distribution of income. The findings underscore the importance of integrating spatial considerations into inequality analysis to better understand the patterns and drivers of regional disparities, thereby informing more effective and equitable policy interventions.

> Geographic income inequality has risen more than 40% between 1980 and 2021.
> – U.S. Department of Commerce (2023)

1 Introduction

The study of spatial, or geographical, disparities is crucial for both scientific and policy-oriented reasons. Scientifically, understanding these disparities allows researchers to uncover patterns and correlations that are vital for advancing knowledge in various fields such as economics (Kanbur, Venables 2005), public health (Deb Nath, Odoi 2024), and environmental science (Venter et al. 2023). From a policy perspective, recognizing and addressing geographical disparities is essential for promoting social equity and economic development. A prime example is the European Union's Cohesion Policy, where the reduction of spatial disparities between member regions takes center stage (Widuto 2019). Additionally, understanding spatial disparities can inform urban planning and environmental policies to create more sustainable and resilient communities.

By bridging the gap between scientific research and policy implementation, the study of geographical disparities helps in crafting evidence-based strategies that promote balanced regional development, reduce inequalities, and improve the overall quality of life. Ultimately, this interdisciplinary approach fosters a deeper understanding of the complex dynamics at play and supports the creation of more inclusive and effective policies.

The field of spatial data science (Rey et al. 2023) provides tools to visualize and analyze spatial inequality. Methods such as local spatial autocorrelation analysis (Anselin 1995), spatial distribution dynamics (Rey 2015), and regionalization (Wei et al. 2020) offer robust frameworks to identify patterns, relationships, and variations across geographical contexts. Thus, spatial data science is well-positioned to support such interdisciplinary research. The goal of this paper is to introduce social scientists to the concepts and measurement of spatial inequality. The emphasis is on adopting a computationally focused

and reproducible treatment that would allow researchers to apply the methods introduced here to their own investigations.

The focus is on a case study of regional income inequality in Mexico, which provides a compelling example of spatial disparities and their implications for both scientific inquiry and policy-making. Mexico's diverse regional landscape, marked by significant economic, social, and cultural differences, offers a rich context for analyzing how inequality manifests across geographical spaces. By examining income inequality at the regional level, this case study highlights the spatial patterns and clusters of economic disparities, providing insight into the underlying processes that drive inequality.

The paper proceeds as follows. We first develop a conceptual understanding of the measurement of spatial inequality. Next, we describe the computational environment employed to analyze spatial disparities. The specific case study is then introduced. We then discuss different a-spatial approaches towards measuring inequality, followed by a detailed exploration of spatially explicit approaches for measuring geographical disparities. The paper concludes with the identification of future research areas in the field of spatial inequality.

2 Inequality Concepts

The growing concern with inequality brings to mind Peter Drucker's often-cited principle, "You can't improve what you don't measure." Before we can address the technical challenges of measuring spatial inequality, it is essential to first grapple with the conceptual issues surrounding what we are measuring.

It is important to distinguish between terms that frequently appear in the inequality literature: equality (inequality) and equity (inequity). Equality refers to the state of being equal, particularly in status, rights, and opportunities. In economics, this often means distributing resources and opportunities uniformly across all individuals or groups.

Equity, conversely, involves fairness and justice in the distribution of resources and opportunities. It considers individual needs and circumstances, aiming to level the playing field. Therefore, although inequality and inequity are interconnected within the context of social justice (Sen 2004), they are not synonymous.

For instance, an equal distribution of resources, such as uniform per capita expenditure on students, can create inequities by ignoring the challenges faced by students in different contexts, such as urban versus rural districts or advantaged versus disadvantaged neighborhoods (Tine 2017). Conversely, some distributions are intentionally unequal to achieve greater equity. Progressive income tax schemes, where the tax rate increases with income, are a prime example of this approach (Ledić et al. 2023).

A closely related distinction is between equality of outcomes and equality of opportunities. Inequality in outcomes refers to the unequal distribution of income, wealth, and resources among individuals in a society, which can result from factors such as luck, effort, and inherited wealth. In contrast, inequality in opportunities focuses on the unequal access to education, healthcare, and other essential services that enable individuals to achieve their potential, irrespective of their background (Roemer 1998). Studies may differ, then, in whether they measure spatial inequality in outcomes (Khedmati Morasae et al. 2024) or spatial disparities in opportunities (Knaap 2017).

In addition to the distinction between inequality and inequity, and between outcomes versus opportunities, there is much variation in the substantive variable under focus. Income studies dominate the literature quantitatively (Gaubert et al. 2021) and are sometimes contrasted with studies of the inequality of wealth (Suss et al. 2024).[1] More granular studies examine disparities in the sources of income, such as wages, as well as employment rates (Overman, Xu 2022). Outside of economics, topics such as disparities in educational outcomes (Graetz et al. 2020), health outcomes (Khedmati Morasae et al. 2024), voting patterns (Barber, Holbein 2022), among many others, are replete across the social and life sciences.

[1] It is important to keep in mind that income is measured as a flow whereas wealth is a stock. This distinction matters in terms of the way disparities in the two variables are examined. See the discussions in (Saez, Zucman 2016) and (Piketty 2014).

The unit of measurement employed in inequality analysis is also an important consideration. Sala-i-Martin (2006) demonstrates that when using countries as the unit of analysis, the picture that emerges is one of large and static levels of international inequality. However, when the analysis used countries weighted by their populations, the view is one of declining inequality over time. Later, we shall see that this choice of weighting versus non-weighting of the units is an important issue in spatial inequality measurement.

There is much variation across the inequality literature in the unit of measure. Studies of personal income inequality often focus on data recorded for individuals (Piketty, Saez 2003). Other studies take the household or family as the unit of analysis (Brandolini, Smeeding 2011). In both cases, the focus is on income inequality across people. This is an essential point of departure for our study of spatial inequality, which is where the unit of measurement is a geographical area. In other words, in spatial inequality analysis, the focus is inequality across places.

Still further, some studies examine the spatial distribution of personal income inequality (Partridge et al. 1998, Frank 2009)—that is, how inequality between individuals within a state varies across states. In the spatially oriented inequality studies, the geographical unit of analysis can range from countries (Milanović 2018), to intra-national regions (Ganong, Shoag 2017), to cities (U.S. Department of Commerce 2023, Sarkar et al. 2024), and down to neighborhoods (Nijman, Wei 2020).

A final inequality concept we need to consider is the role of time. One question is the time unit. Is income measured per person, per year, or is some life-time earnings, or permanent income (Hall 1978) measure employed? A second question pertains to whether the study of inequality is a snapshot at one point in time or focuses on the dynamics of income distribution.

Addressing all these issues is beyond the scope of any one study. We raise them here in order to situate the study of spatial inequality in a much broader context. For this paper, we will hone in on the question of measuring spatial income inequality at the sub-national scale.

3 Computational Environment

In the following section, we present the packages and computational environment used. The narrative following code cells explains the computational concepts.

3.1 Packages

[1]:
```
import inequality as ineq                                                    ①
import numpy as np                                                           ②
import pandas as pd
import geopandas as gpd
import libpysal as lps
import seaborn as sns
import matplotlib.pyplot as plt
import watermark                                                             ③
%load_ext watermark
%watermark -a "Sergio Rey" -u -n -t -v \
    -p numpy,pandas,scipy,matplotlib,inequality,seaborn,libpysal
```

Notes:
① We alias the package `inequality` as `ineq`
② We do the same for `numpy`
③ `watermark` reports the current versions of our packages to support reproducibility

[1]:
```
Author: Sergio Rey

Last updated: Mon Dec 16 2024 09:17:39

Python implementation: CPython
Python version       : 3.10.16
IPython version      : 8.30.0
```

```
numpy     : 2.2.0
pandas    : 2.2.3
scipy     : 1.14.1
matplotlib: 3.9.4
inequality: 1.0.2.dev78+gbb07357
seaborn   : 0.13.2
libpysal  : 4.12.1
```

Our analysis of spatial inequality utilizes packages from the Python Spatial Analysis Library (`pysal`) (Rey et al. 2022), together with `geopandas` (Jordahl et al. 2019) and `seaborn` (Waskom 2021). The main package focusing on measuring spatial disparities is `inequality`, which implements analytics for measuring a-spatial inequality and spatially explicit inequality measures. Also, from `pysal`, we will depend on the `libpysal` package for constructing spatial weights that are central to the analysis of spatial inequality. `geopandas` provides for spatial data processing and producing maps of the spatial distribution of income, while we adopt `seaborn` for constructing a-spatial graphical views of inequality.

For purposes of reproducibility, we include the `watermark` package which reports the version numbers of each of the packages we use in our analysis.

4 Data

To illustrate the core concepts in spatial inequality measurement, we will rely on a data set for the states in Mexico (Rey, Sastré-Gutiérrez 2010). The variable of interest is state per capita gross domestic product (pcgdp) measured in 2000 USD measured for each decade from 1940-2000 for each of 32 areas consisting of the 31 federal states of Mexico plus Mexico City. This data-set is included as an example data-set in `libpysal`.

[2]:
```
lps.examples.explain("mexico")
```

[2]:
```
mexico
======

Decennial per capita incomes of Mexican states 1940-2000
--------------------------------------------------------

* mexico.csv: attribute data. (n=32, k=13)
* mexico.gal: spatial weights in GAL format.
* mexicojoin.shp: Polygon shapefile. (n=32)

Data used in Rey, S.J. and M.L. Sastre Gutierrez. (2010) "Interregional inequality
dynamics in Mexico." Spatial Economic Analysis, 5: 277-298.
```

In addition to the income data contained in the `mexico.csv` file, there are two additional files available in this example: `mexico.gal` which stores information about the contiguity relationships between the states, and `mexicojoin.shp` which is a shapefile.

The following code block produces Figure 1 which lists the locations of the 32 Mexican states.

[3]:
```
# Create a map to provide
# context for the subsequent analysis

# We use the library mapclassify for choropleth classifications
import mapclassify

# set the path to our shapefile and read the file into a geodatafame
pth = lps.examples.get_path("mexicojoin.shp")
gdf = gpd.read_file(pth)

# we will use greedy from mapclassify
# states to ensure contiguous states are of a different color
sgdf = gdf.sort_values(by='NAME')
sgdf.reset_index(inplace=True)
sgdf['label'] = range(1, 33)
sgdf['greedy'] = mapclassify.greedy(sgdf)
```

```
# set the font size and position
font_size = 9
outside = [9, 29]
oc = [(-103, 17.5), (-95, 22.5)]
oe = [(-102.55, 17.49),(-95.5, 22.1)]
oinfo = zip(outside, oc)

# we will use LineStrings from shapely
# for call-outs to state names
from shapely.geometry import LineString

# plot the map
import matplotlib.pyplot as plt
sgdf['centroid'] = sgdf.centroid
ax = sgdf.plot(
    figsize=(8, 12),
    column="greedy",
    categorical=True,
    cmap="Set3",
    #legend=True,
    edgecolor="w",
)

# build the table of state names
table = []
for idx, row in sgdf.iterrows():
    centroid = row['centroid']
    table.append(f'{idx+1:2d} {row["NAME"]}')
    if idx+1 not in outside:
        ax.text(centroid.x, centroid.y, str(idx+1), ha='center',
                va='center', fontsize=font_size, color='black')

# add the call-outs and number the polygons
i = 0
for out in oinfo:
    idx, coords = out
    ax.text(coords[0], coords[1], str(idx), ha='center',
            va='center', fontsize=font_size, color='black')
    start_point = coords
    end_point = sgdf.centroid[idx-1]

    # Create a LineString object
    start_point = oe[i]
    line = LineString([start_point, end_point])

    # Create a GeoSeries for the line
    line_gdf = gpd.GeoSeries([line])

    # Plot the line
    line_gdf.plot(ax=ax, color='red', linewidth=2)
    i+=1

for i, label in enumerate(table):
    if i < 16:
        ax.text(-120, 20-i*1, label, ha='left',
                va='center', fontsize=font_size, color='black');
    else:
        ax.text(-110, 20-(i-16)*1, label, ha='left',
                va='center', fontsize=font_size, color='black');
ax.set_axis_off()
```

[3]:
```
Output in Figure 1
```

We can read the `mexico.csv` file using `pandas` to create a `DataFrame` that will hold the attributes of interest. This dataframe has the shape `(32,13)` indicating there are 32 observations on 13 variables. The 13 variables include the per capita gross domestic product for each decade, for example `pcgdp1990` for 1990, together with the state name, and five variables that define different regionalization schemes for the country. We will return to these regional variables in a later section.

1 Aguascalientes
2 Baja California Norte
3 Baja California Sur
4 Campeche
5 Chiapas
6 Chihuahua
7 Coahuila De Zaragoza
8 Colima
9 Distrito Federal
10 Durango
11 Guanajuato
12 Guerrero
13 Hidalgo
14 Jalisco
15 Mexico
16 Michoacan de Ocampo
17 Morelos
18 Nayarit
19 Nuevo Leon
20 Oaxaca
21 Puebla
22 Queretaro de Arteaga
23 Quintana Roo
24 San Luis Potosi
25 Sinaloa
26 Sonora
27 Tabasco
28 Tamaulipas
29 Tlaxcala
30 Veracruz-Llave
31 Yucatan
32 Zacatecas

Figure 1: States of Mexico

[4]:
```
# Create a DataFrame from a csv file
pth = lps.examples.get_path("mexico.csv")
df = pd.read_csv(pth)
print(f"Shape of dataframe: {df.shape}")
print(f"First 5 rows of dataframe:\n {df.head()}")
df.head()
print(f"\nVariables: {df.columns}")
```

[4]:
```
Shape of dataframe: (32, 13)
First 5 rows of dataframe:
                State  pcgdp1940  pcgdp1950  pcgdp1960  pcgdp1970  pcgdp1980
0       Aguascalientes    10384.0     6234.0     8714.0    16078.0    21022.0
1     Baja California     22361.0    20977.0    17865.0    25321.0    29283.0
2  Baja California Sur     9573.0    16013.0    16707.0    24384.0    29038.0
3            Campeche     3758.0     4929.0     5925.0    10274.0    12166.0
4             Chiapas     2934.0     4138.0     5280.0     7015.0    16200.0

   pcgdp1990  pcgdp2000  hanson03  hanson98  esquivel99  inegi  inegi2
0    20787.0    27782.0       2.0       2.0         3.0    4.0     4.0
1    26839.0    29855.0       1.0       1.0         5.0    1.0     1.0
2    25842.0    26103.0       2.0       2.0         6.0    1.0     1.0
3    51123.0    36163.0       6.0       5.0         4.0    5.0     5.0
4     8637.0     8684.0       5.0       5.0         7.0    5.0     5.0

Variables: Index(['State', 'pcgdp1940', 'pcgdp1950', 'pcgdp1960', 'pcgdp1970',
       'pcgdp1980', 'pcgdp1990', 'pcgdp2000', 'hanson03', 'hanson98',
       'esquivel99', 'inegi', 'inegi2'],
      dtype='object')
```

5 Measuring Spatial Inequality in Mexico

5.1 Visualizing Inequality in Distributions

We begin with different perspectives on the distribution of state incomes in Mexico. There are two different approaches towards visualizing the distribution: one focusing on the geographical distribution and the second on the attribute distribution. From a geographical perspective, Figure 2 shows the spatial distribution of incomes for the last year of the sample (2000).

[5]:
```
# Create a choropleth map of per capita gross
# domestic product in 2000 using quintiles for the classification
pth = lps.examples.get_path("mexicojoin.shp")
gdf = gpd.read_file(pth)
ax = gdf.plot(column="PCGDP2000", k=5, scheme="Quantiles",
              legend=True,
              edgecolor='grey',
              legend_kwds={"loc": "center left",
                           "bbox_to_anchor": (1, 0.5),
                           "fmt": "{:.0f}"})
ax.set_axis_off()
ax.set_title("PC GDP 2000");
```

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

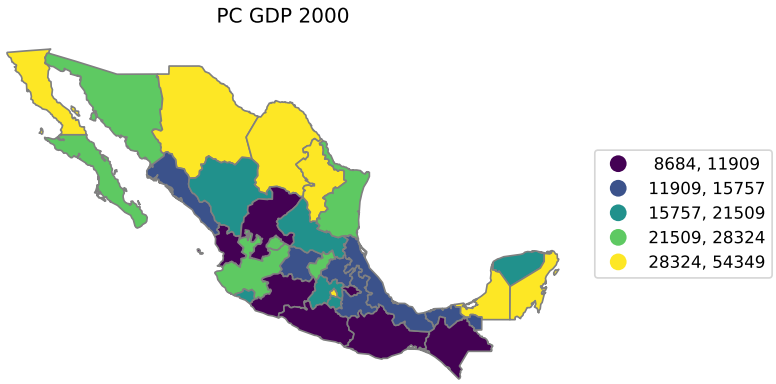


Figure 2: Per Capita Gross Domestic Product by State (Quintiles)

This is a choropleth map using quintiles to classify the incomes. The visual impression is that incomes are not randomly distributed in Mexico, as the states with incomes below the bottom quintile are more concentrated in the south, while in the north, the highest income states dominate. We will be able to make more quantitative evaluation of this spatial pattern later on in this paper.

Figure 3 presents a different perspective on income distribution based on the concept of Pen's Parade, introduced by Dutch economist Jan Pen (Pen 1971). The metaphor uses a Parade to illustrate economic inequality, with each person representing a state in the economy, and their height being proportional to the state's per capita income. The Parade starts with the shortest individuals depicting the poorest states, gradually increasing in height as income rises. At the end of the parade, the tallest individuals represent the most affluent states, highlighting the significant income disparities across states. This visual tool showcases the inequality in income distribution. The parade uses two different scales, with the x-axis showing the ordinal distances between states and the y-axis representing the interval distances in their per capita incomes.

[6]:
```
# Produce the Pen's Parade for 2000
from inequality.pen import pen
f = pen(gdf, 'PCGDP2000', 'NAME')
```

[6]:
```
Output in Figure 3
```

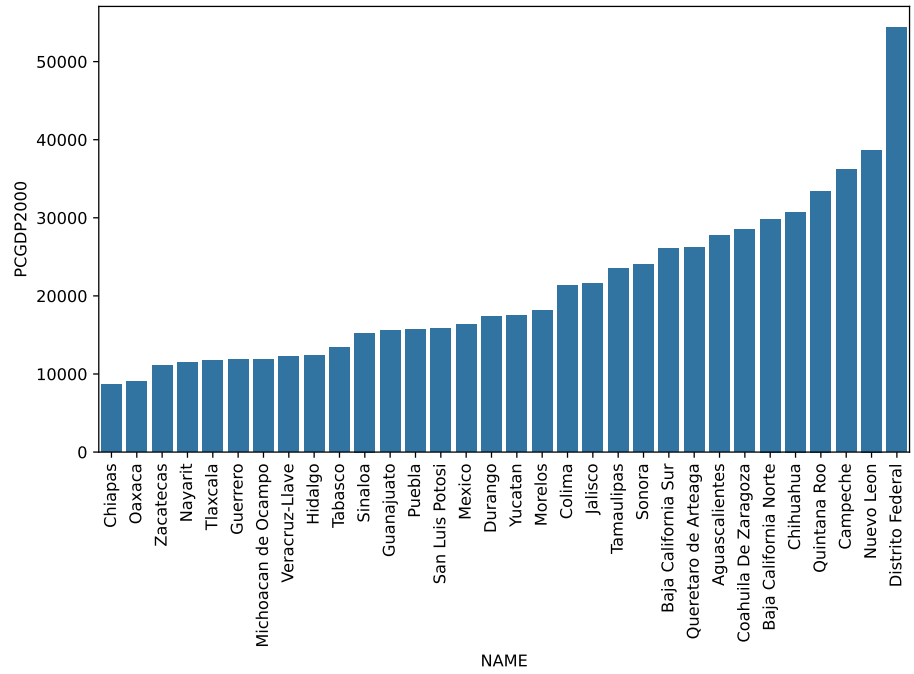


Figure 3: Pen's Parade Per Capita Gross Domestic Product by State 2000

We can start to integrate the spatial and attribute distributions together using a `pengram` from the `pysal-inequality` package:

[7]:
```
# Produce pengram for 2000
from inequality.pen import pengram
f = pengram(gdf, 'PCGDP2000', 'NAME', xticks=False, leg_pos='lower left',
        fmt="{:.0f}")
```

[7]:
```
Output in Figure 4
```

As shown in Figure 4, the `pengram` combines the Pen's Parade alongside the choropleth map. This affords a more granular view of the distribution than those offered by either view in isolation. For example, one of the well-known limitations of a quintile classed map is that the intra-class variation is obscured. In the `pengram`, the intra-class variation now becomes visible through the Pen's Parade, revealing the much larger absolute and relative variance above the upper quintile relative to the other classes.

A second feature of the `pengram` is the ability to query for specific observations. This makes it possible to locate the position of a state in both the Pen's Parade (attribute space) as well as on the map (geographical space). We do this for states occupying the two extremes of the attribute distribution: Chiapas and Distrito Federal in Figure 5. While the two reside in the extremes of the attribute distribution, the high income Distrito Federal is in the center of the geographic distribution while Chiapas is on the southern border of the country. Moreover, although Distrito Federal stands out in the Pen's Parade, its small geographic area makes it difficult to identify on the map without the query functionality of the pengram.

[8]:
```
# Code snippet to produce the pengram
# with query for Chiapas and Distrito Federal
from inequality.pen import pengram
f = pengram(gdf, 'PCGDP2000', 'NAME',  leg_pos='lower left',
            fmt="{:.0f}",
            xticks=False,
            query=['Chiapas', 'Distrito Federal'])
```

[8]:
```
Output in Figure 5
```

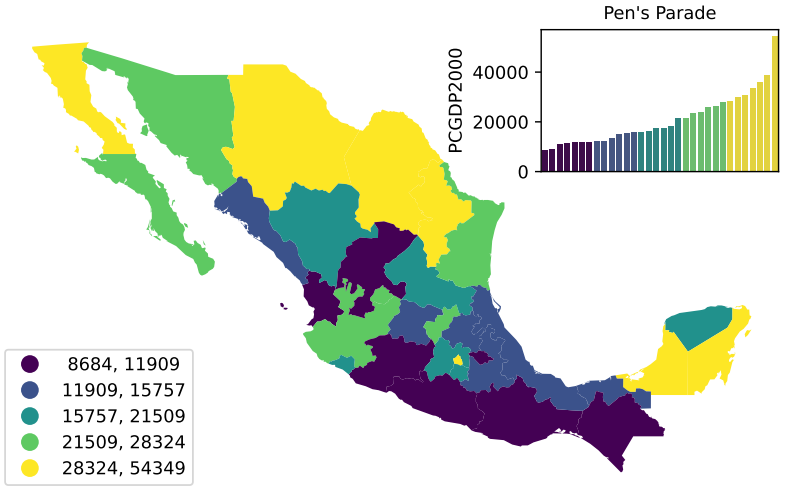


Figure 4: Pengram Per Capita Gross Domestic Product by State 2000

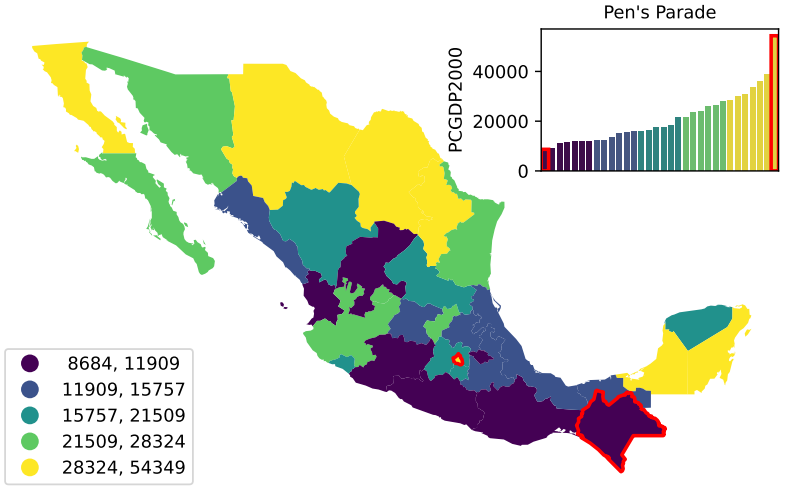


Figure 5: Querying the Pengram of Per Capita Gross Domestic Product by State 2000

Returning to a more granular view of the attribute distribution, Figure 6 combines a histogram of the distribution together with a kernel density estimate, and a rug plot. The kernel density estimate is a smooth curve that represents the probability density function of the data, providing a continuous approximation of the underlying distribution. The rug plot signifies the positions of each state as short ticks on the x-axis. The outlier nature of the Distrito Federal that we saw in the `pengram` is responsible for the positive (right) skew of the density function. There is some evidence of polarization in the distribution with the mode being at the poorest group of states and other, lower, peaks in the middle of the distribution.

[9]:
```
# Produce histogram,
# kernel density and rug plot
years = range(1940, 2010, 10)
yvars = [f'pcgdp{year}' for year in years]
sns.histplot(df[yvars[-1]], kde=True, element="step",
             bins=10, edgecolor="black")
sns.rugplot(df[yvars[-1]], color='blue')
plt.show()
```

[9]:
```
Output in Figure 6
```

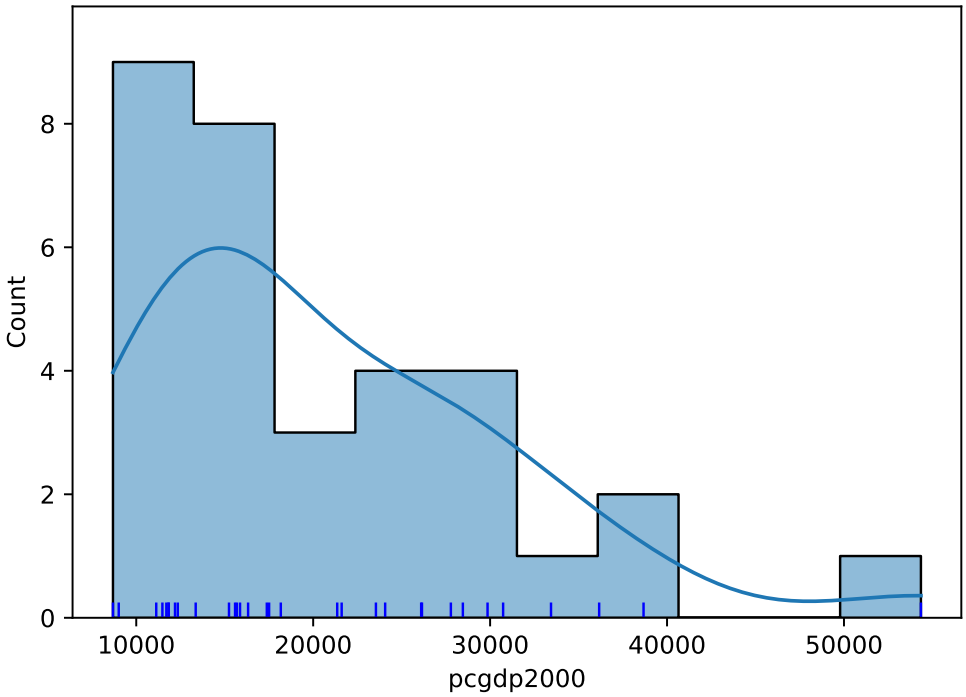


Figure 6: Distribution Per Capita Gross Domestic Product by State 2000 (pcgdp2000), histogram, kernel density, rug plot

Inequality in a distribution is often considered by an inspection of the shares of overall income belonging to units at different locations within the distribution. Here, we must keep in mind, the distinction between the different units under study in the analysis of spatial versus personal income inequality. In spatial inequality analysis, we essentially treat each state as an "individual" and set that individual's level of income to the state's per capita income. The share for the state is then derived as the ratio of its per capita income to the sum of the per capita income of all states.

These shares can be portrayed in a Lorenz curve, shown in Figure 7, which orders the states by their per capita incomes from lowest to highest. Then, against the cumulative proportion of states (x-axis) we plot the cumulative income share on the y-axis. Both scales have limits of $[0, 1]$. In the case of perfect equality, where all state per capita incomes are equal, this plot would be a 45-degree line, the so called line of perfect equality. Any departure from perfect equality will result in states with above average per capita income receiving more than $1/n$ share of $n\bar{y}$, while states with below average per capita incomes will have shares below $1/n$.[2]

[10]:
```
# Produce Lorenz Curve and Schutz Line
from inequality.schutz import Schutz
s = Schutz(gdf, 'PCGDP2000')
s.plot(xlabel='Share of States', ylabel='Share of Per Capita Income')
```

[10]:
```
Output in Figure 7
```

5.2 Measures of Inequality

A large number of measures of inequality exist, and choosing among the rich diversity of inequality measures has been the subject of a vast literature.[3] To help guide that selection, there are five desirable properties of an inequality measure:

[2] States with shares below $1/n$ would also have location quotients of less than one for their relative per capita incomes, where their per capita income was expressed relative to national per capita income.

[3] See Cowell (2011) for an overview of inequality measures.

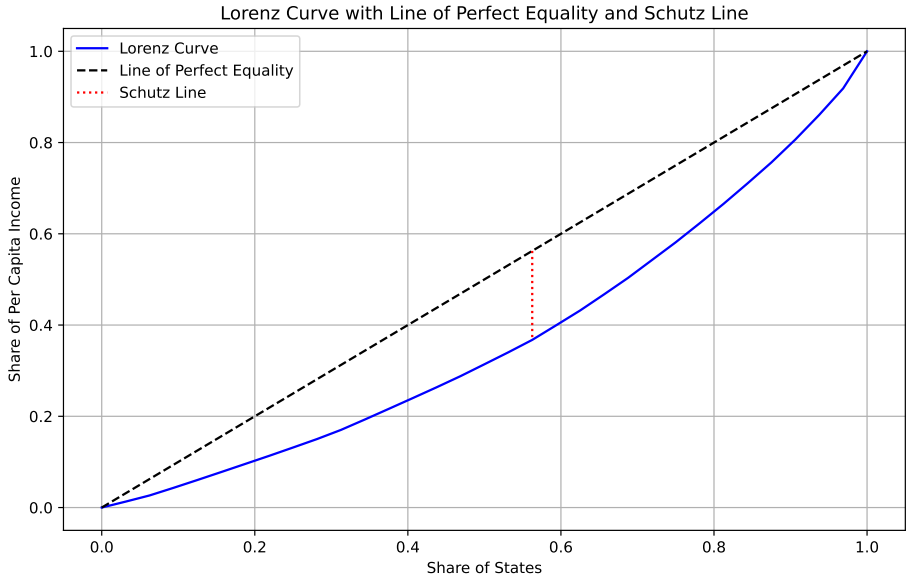


Figure 7: Lorenz Curve and Schutz Line Per Capita Gross Domestic Product by State 2000

1. Symmetry or anonymity
2. Principle of transfers
3. Scale invariance
4. Replication invariance
5. Zero normalization

Symmetry implies that the names of the income receiving unit should be immaterial. That is, if we swap the income of one geographical unit with that of another, the overall inequality measure should not change. The principle of transfers implies that the measure should reflect a reduction in inequality if income is transferred from a richer unit to a poorer unit, as long as the transfer does not reverse their income ranking. Scale invariance means that if all incomes are multiplied by the same constant, inequality remains unchanged. Replication invariance implies that an inequality measure should be unaffected if the population is replicated, meaning that duplicating the entire income distribution does not alter the measure of inequality. Zero normalization means that an inequality measure assigns a value of zero to a perfectly equal income distribution, serving as a baseline where all units have the same income.

Not all of the inequality measures satisfy all of these five properties. For example, the variance is not scale invariant, and both the logarithmic variance and variance of logarithms fail the principle of transfers property. Even for measures that respect these five properties, their use in comparing the levels of inequality between countries at one point in time, or the same country at different points of time, requires the researcher to specify a social welfare function (Atkinson 1970).

We can derive two indices from the Lorenz curve: the Gini coefficient and the Schutz coefficient. The Gini coefficient is an area measurement expressed as a ratio of the area of the lens above the Lorenz curve but below the diagonal to area under the line of perfect equality. The Schutz coefficient is a distance measure defined as the maximum vertical distance between the equality diagonal and the Lorenz curve. Both coefficients measure the extent to which the Lorenz curve departs from perfect equality. Moreover, both indices run from 0, perfect equality, to 1, maximal inequality.

For 2000, the Gini coefficient is 0.258, while the Schutz coefficient is 0.195. As they are both scaled to be within the unit interval, the temptation is to compare the two values. However, this would be misleading as we recall one is a measure of area, the second a measure of distance. Instead, each coefficient is often used to compare across distributions, either over time or space. We can do this for Mexico by asking what has

happened to inequality over time. Along the way, we add a third commonly used indicator of inequality, the coefficient of variation (CV). The coefficient of variation is a relative measure of variation as it is defined as the ratio of the standard deviation to the sample mean.

[11]:
```
# Code snippet to calculate inequality measures for each decade
# and display in a table
yvars = [f'PCGDP{year}' for year in years]
ginis = [ineq.gini.Gini(gdf[yvar]).g for yvar in yvars]
res_df = pd.DataFrame(data=ginis, columns=['Gini'], index=years)
cv = gdf[yvars].std() / gdf[yvars].mean()
res_df['CV'] = cv.values
s = [ineq.schutz.Schutz(gdf, yvar).distance for yvar in yvars]
res_df['Schutz'] = s
res_df['Gini_rank'] = res_df['Gini'].rank()
res_df['CV_rank'] = res_df['CV'].rank()
res_df['Schutz_rank'] = res_df['Schutz'].rank()
res_df
```

[11]:
```
Output in Table 1
```

Table 1 shows that the three indices agree that inequality was lowest in 1980, and the highest in 1940. However, while the Gini and Schutz coefficients agree in their rankings, the inclusion of the CV creates discordance. Part of the discordance reflects the sensitivity of the measures to different parts of the income distribution. The Gini coefficient puts more weight on the middle of the distribution, while the CV is more affected by the right tail of the distribution. The discordance also reflects the property that when the Lorenz curves do not intersect, the CV and Gini would agree on the rankings of inequality. However, Figure 8 shows that there are cases where the Lorenz curves intersect.

The discordance in rankings complicates whether income inequality between states in Mexico has increased or decreased. The answer now depends upon which temporal interval one chooses. Both the Gini and CV agree that inequality has declined since 1940, irrespective of the terminal year selected. However, they disagree on the answer when the question is whether inequality decreased from, say, 1960 to 1970 or between 1990 and 2000 (see Table 1).

[12]:
```
# Plot Lorenz Curves for each decade
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import cumfreq

def lorenz_curve(incomes):
    sorted_incomes = np.sort(incomes)
    cumulative_incomes = np.cumsum(sorted_incomes)
    normalized = cumulative_incomes / cumulative_incomes[-1]
    lorenz_curve = np.insert(normalized, 0, 0)
    n = len(incomes)
    x = np.linspace(0.0, 1.0, n + 1)
    return x, lorenz_curve
```

Table 1: Inequality Index Rankings

	Gini	CV	Schutz	Gini_rank	CV_rank	Schutz_rank
1940	0.353724	0.719858	0.260037	7.0	7.0	7.0
1950	0.296446	0.624611	0.213920	6.0	6.0	6.0
1960	0.253718	0.492447	0.181549	3.0	3.0	3.0
1970	0.255134	0.472039	0.185266	4.0	2.0	4.0
1980	0.245053	0.462657	0.179702	1.0	1.0	1.0
1990	0.251818	0.497729	0.181363	2.0	5.0	2.0
2000	0.258113	0.492565	0.195043	5.0	4.0	5.0

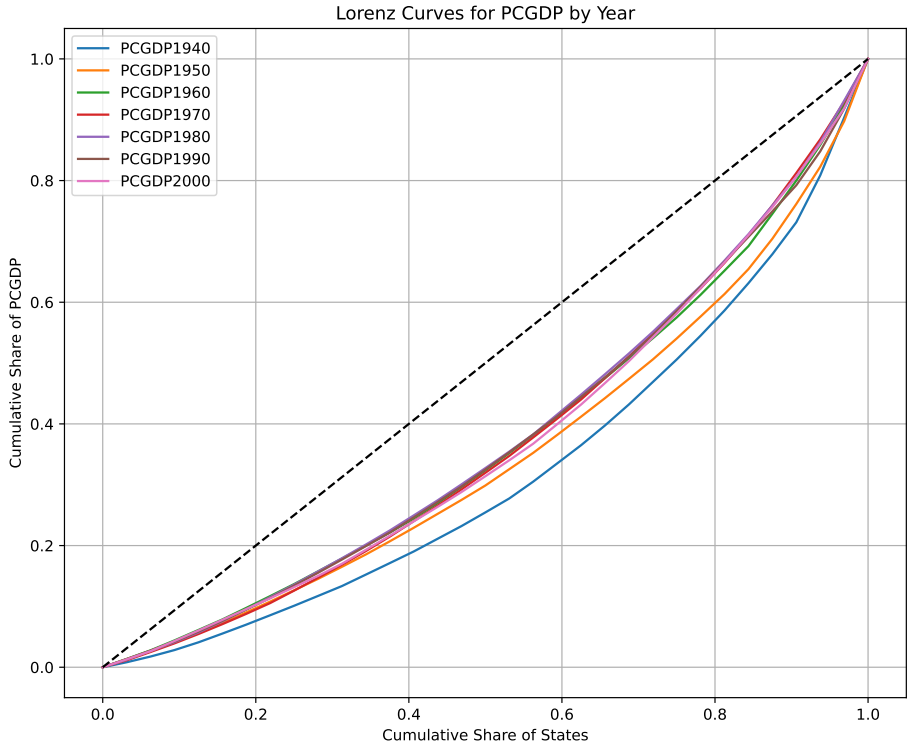


Figure 8: Lorenz Curves by Decade

```
plt.figure(figsize=(10, 8))

for yvar in yvars:
    incomes = gdf[yvar].values
    x, y = lorenz_curve(incomes)
    plt.plot(x, y, label=yvar)

# Plotting the line of equality
plt.plot([0, 1], [0, 1], color='black', linestyle='--')

# Adding titles and labels
plt.title('Lorenz Curves for PCGDP by Year')
plt.xlabel('Cumulative Share of States')
plt.ylabel('Cumulative Share of PCGDP')
plt.legend()
plt.grid(True)
plt.show()
```

[12]:
```
Output in Figure 8
```

In the study of spatial inequality, it is important to note that all the measures mentioned above share a sixth property: *spatial invariance*. This means they are insensitive to the *geographical* distribution of the income values. The spatial invariance is demonstrated in Figure 9, where we compare two spatial distributions, one that is the actual distribution for 2000 and one where we artificially permute the incomes randomly. Despite markedly different spatial patterns, the income distributions summarized in the histograms in the bottom row are identical. Since all of the inequality measures mentioned only consider information about the statistical distribution, each measure will take on the same value whether applied to the spatial distribution on the left or right of the figure.

From a geographical perspective, spatial invariance is not a desirable property in an inequality measure. Let's now turn our attention to spatially explicit measures of inequality.

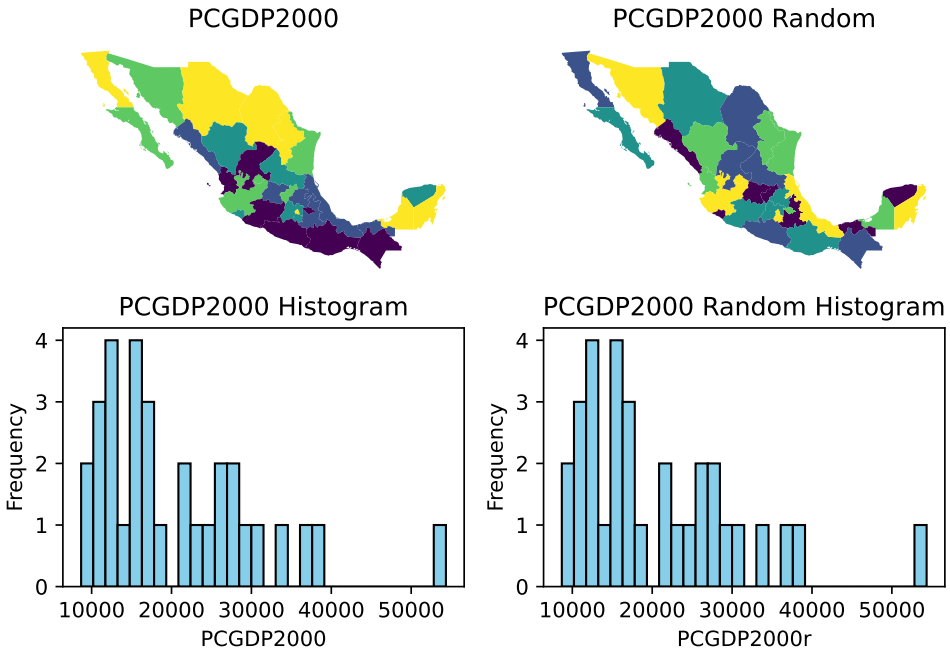


Figure 9: Spatial Invariance of Distributions

[13]:
```
# Generate a 2x2 grid
# Choropleths in first row
# Income histograms in second row
fig, axs = plt.subplots(2, 2)

gdf.plot(column='PCGDP2000', ax=axs[0, 0], scheme='quantiles',
         cmap='viridis')
axs[0, 0].set_title('PCGDP2000')
axs[0, 0].axis('off')


gdf['PCGDP2000r'] = np.random.permutation(gdf.PCGDP2000)

gdf.plot(column='PCGDP2000r', ax=axs[0, 1], scheme='quantiles',
         cmap='viridis')
axs[0, 1].set_title('PCGDP2000 Random')
axs[0, 1].axis('off')

axs[1, 0].hist(gdf['PCGDP2000'], bins=30, color='skyblue',
               edgecolor='black')
axs[1, 0].set_title('PCGDP2000 Histogram')
axs[1, 0].set_xlabel('PCGDP2000')
axs[1, 0].set_ylabel('Frequency')

axs[1, 1].hist(gdf['PCGDP2000r'], bins=30, color='skyblue',
               edgecolor='black')
axs[1, 1].set_title('PCGDP2000 Random Histogram')
axs[1, 1].set_xlabel('PCGDP2000r')
axs[1, 1].set_ylabel('Frequency')

plt.tight_layout()

plt.show()
```

[13]:
```
Output in Figure 9
```

5.3 Putting Space into the Measurement of Inequality

The inability of the inequality measures to capture any of the geographical dimensions of inequality stems from the treatment of the geographical units of measurement as individuals, and the desire to respect the principle of symmetry or anonymity in a classic inequality measure. This enables researchers to draw upon the wealth of knowledge about the properties of classic inequality measures, but at the cost of ignoring geography. In other words, these measures say a lot about inequality in the statistical distribution of incomes but they are silent on the spatial distribution of incomes.

In this section, we discuss the approaches used to integrate the spatial and statistical distributions in the study of spatial inequality. Given that these methods take the geographical distribution into account, they can be said to be *spatially explicit measures of inequality*.

These approaches are special cases of inequality decomposition. However, we highlight a key distinction between those approaches that use the concept of regional groupings to represent the geographical dimensions of inequality, and those that take a more comprehensive approach to introducing geography by considering the pair-wise spatial relationships between observations. We label the former category as regional inequality decomposition approaches, and the latter as spatial inequality decomposition methods.

5.3.1 Regional Inequality Decomposition

A common approach to introducing geography in the measurement of inequality leverages the fact that certain inequality measures can be *decomposed* if the individual income receiving units are placed into a set of mutually exclusive and exhaustive groups. The decomposition then identifies the overall inequality that is due to inequality within groups and between the groups.

To see how this works, we focus on the Theil inequality index (Theil 1967). The Theil inequality index is a measure of economic inequality derived from information theory, capturing the extent to which a distribution deviates from perfect equality. It is sensitive to differences across the distribution and can be decomposed into "within-group" and "between-group" components to analyze inequality at different levels.

More formally, let y_i be the income of unit i, and s_i represent the income share of unit i such that $s_i = \frac{y_i}{\sum_i y_i}$ and $\sum_i s_i = 1$. Then, consider the distribution of the shares. When all units have the same income $s_i = s_j = 1/n$. Then the entropy of the shares given as

$$H(y) = \sum_{i=1}^{n} s_i \ln \frac{1}{s_i}$$

will be maximized at $\ln n$. In the case of extreme inequality, all but one unit have $s_i = 0$ and a single unit has all income $s_j = 1$, and $H(y) = 0$. Thus, the entropy function can be viewed as an indicator of *income equality*.

To generate an indicator of *income inequality*, we can contrast an observed distribution's equality against the maximum:

$$T(y) = \ln n - H(y) = \ln n - \sum_{i=1}^{n} s_i \ln \frac{1}{s_i} = \sum_{i=1}^{n} s_i \ln n s_i.$$

This can be viewed as a weighted average of the logarithmic deviations of the shares, with the weights defined as the shares. The logarithmic deviation of share i from perfect equality is $\ln \frac{s_i}{1/n} = \ln s_i n$.

Alternatively, the Theil index can be defined using relative incomes:

$$T = \frac{1}{n} \sum_i \frac{y_i}{\mu} \ln \frac{y_i}{\mu}$$

where y_i is the income of unit i and $\mu = \frac{1}{n} \sum_i y_i$.

Decomposition of the overall T measure requires assigning each unit to exactly one of G sets $S_1, S_2, \ldots, S_G$, with the size of each set given as n_g so that:

$$\sum_{g=1}^{G} n_g = n.$$

Given this, we have:

$$Y_g = \sum_{i \in S_g} y_i \; g = 1, \ldots, G$$

Defining $\omega_g = \frac{n_g}{n} \frac{\bar{Y}_g}{\mu}$, the Theil index can be rewritten as:

$$T = \sum_{g=1}^{G} \omega_g T_g + \sum_{g=1}^{G} \omega_g \ln \frac{\bar{Y}_g}{\mu}. \tag{1}$$

The first term is the within group inequality defined as a weighted average of the inequality within each group with the weights equal to the group's share of overall income, with:

$$T_g = \frac{1}{n_g} \sum_{i \in S_g} \frac{y_i}{\bar{Y}_g} \ln \frac{y_i}{\bar{Y}_g}.$$

The second term in Equation 1 is the between group inequality component, measuring the inequality that would exist if within each group there was no inequality (i.e., all members of the same set have the same income).

Decomposition of inequality had been widely applied in economics to study inequality between occupational groups, sexes, and races, where individuals would be placed into the mutually exclusive groups and overall individual inequality decomposed into that due to the differences between and within the groups. It was a short jump to adopt this to spatial inequality by using regions to define the groups, with individual units (in our case, states) being assigned to one and only one region.[4]

We will illustrate this for the case of Mexican states using a regional partition due to Hanson (1996) as shown in Figure 10. This regionalization scheme consists of 5 regions, with the size of the regions ranging from 2 states to 10 states.

We can apply the `theil` module from `pysal-inequality` to calculate the value of the overall level of inequality as measured by the global Theil index, and its decomposition into the between region inequality and within region inequality components:

[14]:
```
# Produce 2x2 Grid
# First row: spatial distribution and random distribution
# Second row: another random distribution and density of
# polarization values
fig, axs = plt.subplots(2, 2)
from inequality.theil import TheilDSim
np.random.seed(12345)

gdf.plot('PCGDP2000', ax=axs[0, 0], scheme='quantiles',
         cmap='viridis')
axs[0, 0].set_title('PCGDP2000')
axs[0, 0].axis('off')

# Extract the per capita GDP and regimes
income = gdf['PCGDP2000']
regimes = gdf['HANSON98']

res = TheilDSim(income, regimes, 999)

gdf['PCGDP2000r'] = np.random.permutation(gdf.PCGDP2000)
gdf.plot(column='PCGDP2000r', ax=axs[0, 1], scheme='quantiles',
         cmap='viridis')
axs[0, 1].set_title('PCGDP2000 Random')
```

[4] One of the earliest applications of decomposition for regional inequality analysis is Theil (1967).

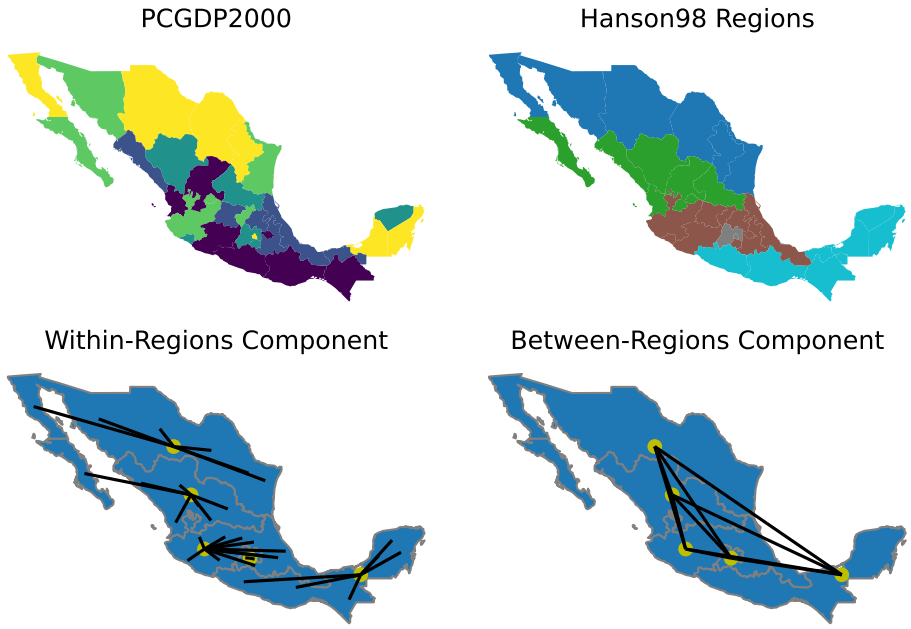


Figure 10: Regional Inequality Decomposition

```
axs[0, 1].axis('off')



gdf['PCGDP2000r'] = np.random.permutation(gdf.PCGDP2000)
gdf.plot(column='PCGDP2000r', ax=axs[1, 0], scheme='quantiles',
         cmap='viridis')
axs[1, 0].set_title('PCGDP2000 Random')
axs[1, 0].axis('off')

msg=f'Spatial polarization: {res.bg[0][0]/res.T:.3f}'
msg=f'{msg}, pseudo p-value: {res.bg_pvalue[0]}'
print(msg)
realizations = np.array([t.bg/t.T for t in res.results])
print(f'Ho mean: {realizations.mean():.3f}')

kde = sns.kdeplot(realizations, fill=False, color='blue', ax=axs[1,1])
x, y = kde.get_lines()[0].get_data()
plt.legend([], [], frameon=False)
# Fill the area to the right of the specified value
plt.fill_between(x, y, where=(x >= realizations[0]),
                 interpolate=True, color='red', alpha=0.5)

# Add vertical line at the specified value
plt.axvline(x=realizations[0], color='red', linestyle='--')
plt.xlabel("Spatial Polarization")
print(f'Theil: {res.T}')
print(f'KB p-value: {(realizations >= realizations[0]).sum()/1000}')
```

[14]:
```
Spatial polarization: 0.341, pseudo p-value: 0.036
Ho mean: 0.139
Theil: 0.10660832349588023
KB p-value: 0.036

Graphical output in Figure 11
```

The overall level of regional inequality is 0.106. The between region component stands at 0.036, while the within region element is 0.070. In relative terms, inequality between the regions in Mexico accounts for 34 percent of state income inequality, while inequality between states from the same region is the larger share presenting 72 percent of spatial inequality.

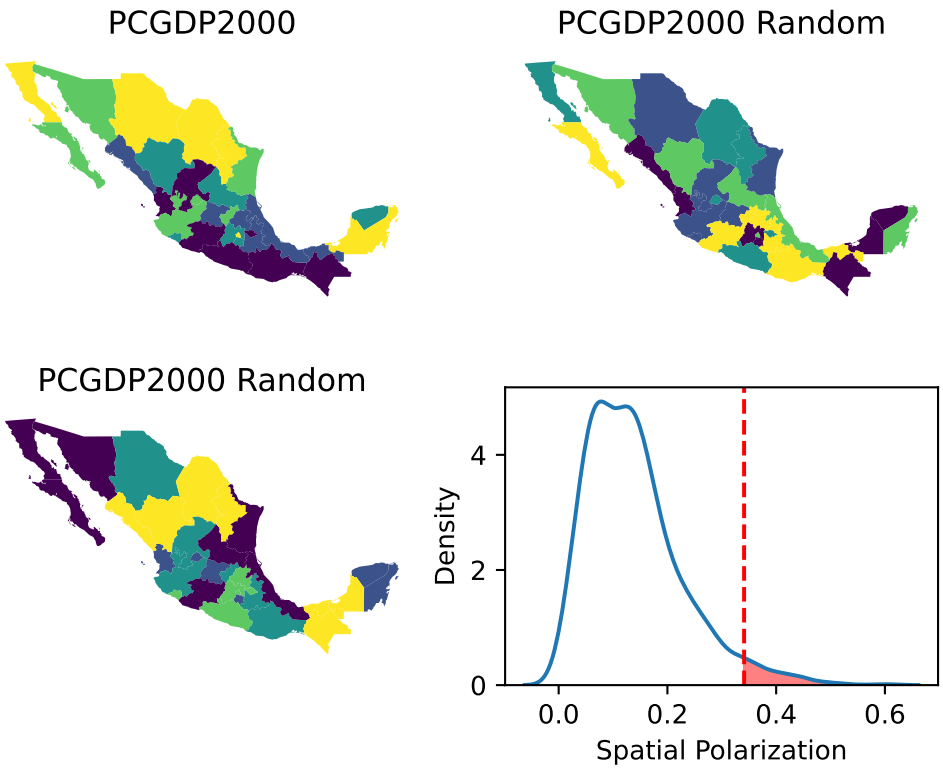


Figure 11: Spatial Polarization

The ratio of between to within region inequality has been suggested as a measure of *spatial polarization* (Zhang, Kanbur 2001). Since the two components sum to overall inequality, we can re-express the measure of spatial polarization as the ratio of between region to overall inequality. Thus, the level of spatial polarization of state incomes stood at 34 percent in 2000 in Mexico.

A relevant question here is whether this level of spatial polarization in 2000 should be considered high or not? Another way to express this, is to ask if the polarization is higher than we would expect if state incomes were randomly distributed in space in Mexico in 2000.

We can answer this question by developing counter-factual spatial distributions that reflect a null hypothesis of spatially random income distributions. Given the n observations, in our case states, there are $n!$ permutations of incomes that are equally likely. In Figure 11, the maps in the top-right and lower-left are two such realizations, where the observed incomes from 2000 (top-left) have been randomly reassigned to states.

Since it is not feasible to generate all 32! maps in our case,[5] we take a sample of 999 such maps from the distribution of permutations. For each of these maps we calculate the spatial polarization measure, and collect all measures to develop a reference distribution for our index under the null of spatially random income distributions. We evaluate our observed spatial polarization index against this distribution and derive a pseudo p-value as the number of counterfactual distributions that generate polarization levels as large as the observed value over the number of permutations plus one.

The reference distribution for the polarization index is reported in the bottom right of Figure 11. The area to the right of our observed polarization index of 0.34 is 0.036 of the distribution. By convention, this p-value points to a significant departure from the null, and we would conclude that the level of spatial polarization of incomes in Mexico is significantly different from that expected were incomes generated by a spatially random process.

The regional decomposition is a spatially explicit measure of spatial inequality as it is

[5] $32! = 2.6313084e + 35$.

indeed sensitive to the geographical distribution of the income values. It is important to note, however, that it is the spatial polarization index, and not the overall T that is spatially explicit. Each of the counterfactual spatial distributions used to construct the reference distribution for the spatial polarization index would still have the same attribute distribution, with similar means and variances, and levels of overall inequality.

While the regional inequality decomposition is spatially explicit, it treats geography in an aggregated fashion. There are, in essence, two scales of spatial inequality implied in this decomposition. As displayed in Figure 10, the within-regions component might be considered the "local" measure as it compares incomes belonging to the same region to the mean income of the group. The between-regions inequality component, by contrast, views spatial inequality in a more aggregate fashion considering only the differences in regional means.

A close inspection of the decomposition Equation 1 reveals both the within and between inequality components are functions of a T index. In the former case this is applied to states from the same region, and in the latter case, the T is applied to the means of regional incomes. For the within region component, the actual spatial distribution of the member states within each of their regions is ignored. By the same token, the geographical location of the regions is immaterial to the calculation of the between region component. In other words, once the states have been assigned to regions, geography no longer matters. Thus, we draw a distinction between *regional* inequality decomposition and *spatial* inequality decomposition.

5.3.2 Spatial Inequality Decomposition

To capture these ignored dimensions of the spatial distribution, Rey, Smith (2013) suggested a spatial decomposition of the Gini coefficient. Starting from the Gini coefficient expressed as half the relative mean absolute deviation:[6]

$$Gini = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{|y_i - y_j|}{2n^2\bar{y}} \tag{2}$$

the sum of the absolute deviations can be decomposed as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j| = \sum_{i=1}^{n}\sum_{j=1}^{n}w_{i,j}|y_i - y_j| + (1 - w_{i,j})\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j| \tag{3}$$

where $w_{i,j} = 1$ if states i and j are geographical neighbors, $w_{i,j} = 0$ otherwise. Here we define neighbors as states that share a border.

Substituting Equation 3 into Equation 2 gives:

$$Gini = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w_{i,j}|y_i - y_j|}{2n^2\bar{y}} + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(1 - w_{i,j})|y_i - y_j|}{2n^2\bar{y}} \tag{4}$$

The first term represents a measure of inequality between neighboring states, while the second term captures inequality between "distant" pairs of states. For most spatial configurations, the number of neighboring pairs will be dwarfed by the number of pairs of states that are distant. So while, in the case of spatially clustered income values, the expectation would be for the average difference in incomes to be smaller for neighboring rather than distant states, our measure of spatial clustering here has to take into account the different cardinality of the two sets of pairs. As such, the relevant comparison is if the first term (second term) is smaller (larger) than what could be expected if state incomes were randomly distributed in space.

We apply the Spatial Gini Decomposition to Mexican State incomes in 2000 using `pysal-inequality` in Figure 12. The adjacency graph based on the criterion of Queen neighbors is shown in the top-right figure. An edge defines a pair of neighboring states.[7]

[6] It is sometimes stated that the maximum value of the Gini in this form is 1 (e.g., Wang et al. 2024). This is technically incorrect, as in in this form, the Gini has a range of $[0, (n-1)/n]$. As n grows larger, the upper bound approaches 1. Moreover, because the upper bound is a function of n, care should be taken when using the Gini in this form when comparing distributions of different sizes.

[7] Two units are Queen neighbors if their borders share at least one vertex.

[15]:
```
# Produce 2x2 Grid
# First row: spatial distribution and neighbor graph
# Second row: random distribution and density of
# Spatial Gini values
fig, axs = plt.subplots(2, 2)
from inequality.gini import Gini_Spatial
import libpysal
np.random.seed(12345)


gdf.plot('PCGDP2000', ax=axs[0, 0], scheme='quantiles', cmap='viridis')
axs[0, 0].set_title('PCGDP2000')
axs[0, 0].axis('off')


wq = libpysal.weights.Queen.from_dataframe(gdf)
wq.transform = 'B'
gs2000 = Gini_Spatial(gdf["PCGDP2000"], wq)

income = gdf['PCGDP1940']

gdf.plot(ax=axs[0, 1])
axs[0, 1].set_title('Neighbor Graph')
axs[0, 1].axis('off')


wq.plot(gdf, ax=axs[0,1])

axs[1, 0].set_title('Counterfactual')
axs[1, 0].axis('off')
gdf['PCGDP2000r'] = np.random.permutation(gdf.PCGDP2000)
gdf.plot(column='PCGDP2000r', ax=axs[1,0], scheme='quantiles',
         cmap='viridis')

adsum = gs2000.dtotal
realizations = gs2000.wcgp / adsum
kde = sns.kdeplot(realizations, fill=False, color='blue', ax=axs[1,1])
x, y = kde.get_lines()[0].get_data()
plt.legend([], [], frameon=False)
# Fill the area to the right of the specified value
plt.fill_between(x, y, where=(x >= gs2000.wcg/adsum),
                 interpolate=True, color='red', alpha=0.5)

# Add vertical line at the specified value
plt.axvline(x=gs2000.wcg/adsum, color='red', linestyle='--')
plt.xlabel("Spatial Gini")


plt.tight_layout()
plt.show()
G = gs2000
msg = f'Expected Distant SADS/Total SADS: {G.e_wcg/adsum:.2f}'
msg = f'{msg}, Observed: {G.wcg/adsum:.2f}\n p-value: {G.p_sim}'
print(msg)
```

[15]:
```
Expected Distant SADS/Total SADS: 0.86, Observed: 0.90
 p-value: 0.01

 Graphical output in Figure 12
```

For inference on the spatial Gini, the same computationally based approach that we used in the Theil decomposition is employed. One of the counterfactuals representing a random permutation of the incomes is shown in the lower-left figure. The reference distribution for the spatial Gini index is shown on the bottom right. Here the spatial Gini is expressed as the share of the overall absolute pair-wise differences due to inequality between distant (non-neighbor) pairs of states. The pseudo p-value (0.01) for the observed index is calculated as the area to its right under the distribution.

Under the null, the distant pairs should account for 86 percent of the absolute differences, however, the observed share is much higher at 90 percent. In fact, the p-value

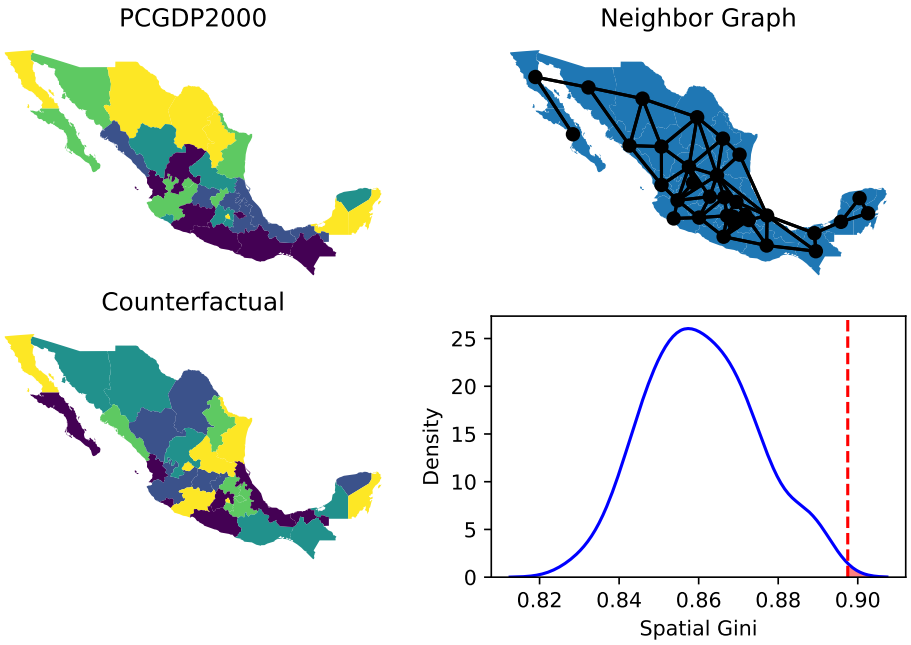


Figure 12: Spatial Inequality Decomposition

indicates that none of the counterfactual spatial distributions of income generated a spatial Gini as large as the one observed. In other words, the inequality between distant pairs of states is larger than the inequality between neighboring states. This pairwise orientation of the spatial Gini decomposition offers a useful complement over the regional decomposition of the Theil approach, as it introduces a more spatially explicit view of the income distribution that demonstrates how spatial autocorrelation affects overall inequality across the states.[8]

Spatial autocorrelation refers to the degree to which a spatial variable is correlated with itself across geographical space. It measures whether similar or dissimilar values of a variable tend to cluster together spatially. Positive spatial autocorrelation indicates that similar values (e.g., areas with high incomes) are located near each other, while negative spatial autocorrelation suggests that dissimilar values (e.g., areas with high and low incomes) are spatially proximate. Spatial autocorrelation is crucial in studying spatial inequality because it reveals the extent to which inequality is spatially patterned, reflecting processes like segregation, clustering of poverty or wealth, and the spatial diffusion of economic opportunities or disadvantages.

Identifying and quantifying spatial autocorrelation allows researchers to understand the spatial dimensions of inequality, assess the effectiveness of place-based policies, and model spatial processes more accurately (Anselin 1995). For example, regions with high positive spatial autocorrelation of income inequality may require targeted regional policies to address concentrated disadvantage. Without considering spatial autocorrelation, analyses of inequality risk overlooking critical spatial dependencies that shape social and economic outcomes. Moreover, the presence of spatial autocorrelation implies that inference on inequality measures that rely on the assumption of random sampling may be misleading. This is because the autocorrelation violates the assumption of independence underlying the random sampling assumption.

While the spatial Gini decomposition does take spatial autocorrelation into account, it does not allow for the exact additivity of within-group and between-group inequality components. This occurs because the Gini index depends on the degree of overlap in incomes among states from different regions. Any overlap complicates the decomposition, as part of the total inequality arises from the overlap itself, which cannot be clearly attributed to either within-group or between-group inequality. As a result, the decomposition requires

[8] For a recent extension of the spatial Gini decomposition see Panzera, Postiglione (2020).

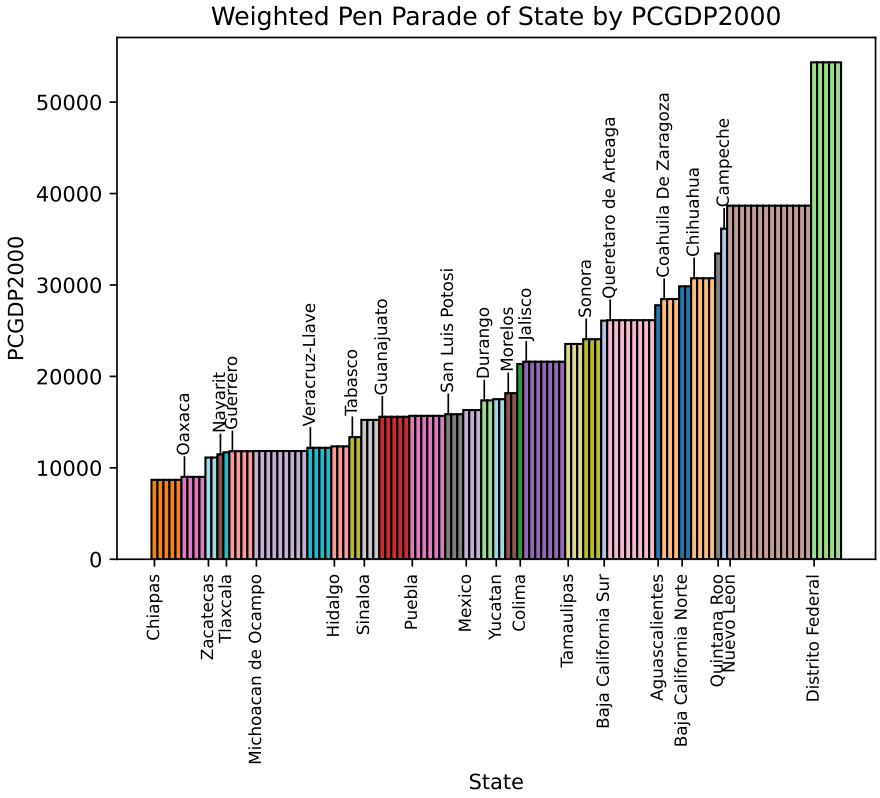


Figure 13: Populated Weighted Pen Parade

a residual term to account for the overlap. Since this residual term is difficult to interpret, researchers rarely use the Gini index for regional inequality decomposition.

5.4 Weighted or Unweighted Inequality: Places versus People

A final issue we explore is the question of whether the measure of spatial inequality should take into account the population sizes of the enumeration units. This was briefly mentioned earlier in the context of measuring international inequality. In the regional literature, a debate rages as to whether population weighted or unweighted approaches should be used to measure spatial inequality (Gluschenko 2018) .

To frame the debate, it is helpful to consider three different concepts of inequality at the international scale suggested by Milanović (2007). Here we adapt them to the question of measuring spatial inequality at the sub-national scale. Concept 1 is unweighted spatial inequality, where each state is one unit of measurement, and we use its per capita income irrespective of the state's population. In addition to being the dominant approach in regional inequality analysis, this concept is at the core of the literature on regional convergence (Rey, Montouri 1999) where the focus is on whether the incomes of poor and rich states in a system are coming together or growing apart over time.

Concept 2 takes into account the population of the individual states, recognizing that a state like Nuevo Leon with a population of 8.6 million in 2000 having its average income increase by 10,000 USD is likely to have a larger impact than is seeing the per capita income of Colima, with a population of 500,000 change by the same amount. So here, the per capita incomes of the states are now weighted by the population of the states. This is weighted spatial inequality.

Finally, Concept 3, measures inequality between all the individuals in the country. Here we would require information on the individuals both in terms of their incomes and their state of residence. Were it available, such data would allow us to measure personal income inequality. However, in the literature on regional inequality, such data is scarce, and so this concept is not operationalized.

This means the debate in the regional literature is between proponents of Concept 1 and Concept 2.

Thus far in this paper, we have adopted the Concept 1 definition of spatial inequality, that is, unweighted spatial inequality. We can contrast this with the perspective offered by Concept 2, by developing the Weighted Pen Parade under concept 2 shown in Figure 13 and comparing it with the Unweighted Pen Parade from Figure 3. The number of bars for each state in the Weighted Pen Parade is proportional to the population of that state. The logic behind the weighted spatial inequality is that rather than having one representative observation (individual) from the state with their level of income equal to the state per capita income, now there are a proportional number of individuals from each state having their level of income equal to the state's per capita income.[9]

A fundamental issue with this approach is that it assumes that inequality among individuals within the state is zero, since all members have the same level of income. This also implies that for two states that have different levels of per capita income, the poorest members of the richer state will be richer than the richest member of the poorer state. This is at odds with empirical reality.

6 Conclusion

Spatial inequality continues to attract the attention of researchers and policy makers alike. This paper has presented the key methodological approaches available to researchers interested in analyzing spatial disparities. By highlighting the challenges posed in adapting classical inequality measures to the question of spatial inequality, the paper draws important distinctions between different types of inequality decomposition approaches based on their treatment of space. These methods reveal the spatial dimensions of inequality, particularly the extent to which income disparities are geographically clustered.

While the focus has been on regional inequality at the sub-national scale, these methods have a broad scope that researchers can apply at different spatial scales, from analyzing geographical disparities at the international level (Redding, Venables 2004), inter-regional (Bathelt et al. 2024), as well as intra-urban scale (OECD 2018). This scope is vital, as the mechanisms of inequality can differ at each scale. For example, trade policy likely plays a more significant role on an international scale. At the same time, industrial restructuring is more influential on an inter-regional scale, and residential sorting is operative on an intra-urban scale. Furthermore, the resulting patterns of spatial inequality may also vary across these scales. The work by Ganong, Shoag (2017) shows a substantial divergence of incomes across states but reports a more mixed pattern when measuring convergence using labor market areas.

Linking the spatial inequality measures to potential policy interventions offers some interesting possibilities. For example, the distinction between inter-regional (between-region) and intra-regional (within-region) inequality in the decomposition may help researchers and policymakers design such interventions in a tailored fashion. If between-region inequality dominates, there may be a strong case for place-based policies. Conversely, if the within-region inequality component accounts for the majority share, more nationally oriented policies, such as federal tax laws, may be more appropriate.

In future research, it would be beneficial to investigate the relationship between spatial inequality and place mobility. Place mobility refers to the economic trajectory of a location within the national context, similar to the concept of inter-generational income mobility (Rey, Casimiro Vieyra 2023). Exploring the link between place mobility and spatial polarization is crucial, especially how the movement of states within the income distribution impacts broader regional inequalities. Moreover, gaining insight into the interaction between place mobility and "place-based policies" could help in designing more effective regional development strategies that reduce inequality while fostering economic mobility.

[9] A referee suggested that if data on individual incomes were available one could use a pentagram for representing within region/state/city inequality to complement Figure 13. In some cases, such as the Netherlands, access to micro level data may be possible under strict conditions.

References

Anselin L (1995) Local indicators of spatial association-LISA. *Geographical Analysis* 27: 93–115. CrossRef

Atkinson AB (1970) On the measurement of inequality. *Journal of Economic Theory* 2: 244–263. CrossRef

Barber M, Holbein JB (2022) 400 million voting records show profound racial and geographic disparities in voter turnout in the United States. *PLoS ONE* 17: e0268134. CrossRef

Bathelt H, Buchholz M, Storper M (2024) The nature, causes, and consequences of inter-regional inequality. *Journal of Economic Geography* 24: 353–374. CrossRef

Brandolini A, Smeeding TM (2011) *Income Inequality in Richer and OECD Countries*. Oxford University Press. CrossRef

Cowell FA (2011) *Measuring Inequality* (3rd ed.). LSE Perspectives in Economic Analysis. Oxford University Press, Oxford

Deb Nath N, Odoi A (2024) Geographic disparities and temporal changes of diabetes-related mortality risks in Florida: A retrospective study. *PeerJ* 12: e17408. CrossRef

Frank MW (2009) Inequality and growth in the United States: Evidence from a new state-level panel of income inequality measures. *Economic Inquiry* 47: 55–68. CrossRef

Ganong P, Shoag D (2017) Why has regional income convergence in the U.S. declined? *Journal of Urban Economics* 102: 76–90. CrossRef

Gaubert C, Kline P, Vergara D, Yagan D (2021) Trends in US Spatial Inequality: Concentrating Affluence and a Democratization of Poverty. *AEA Papers and Proceedings* 111: 520–525. CrossRef

Gluschenko K (2018) Measuring regional inequality: To weight or not to weight? *Spatial Economic Analysis* 13: 36–59. CrossRef

Graetz N, Woyczynski L, Wilson KF, Hall JB, Abate KH, Abd-Allah F, Adebayo OM, Adekanmbi V, Afshari M, Ajumobi O, Akinyemiju T, Alahdab F, Al-Aly Z, Rabanal JEA, Alijanzadeh M, Alipour V, Altirkawi K, Amiresmaili M, Anber NH, Andrei CL, Anjomshoa M, Antonio CAT, Arabloo J, Aremu O, Aryal KK, Asadi-Aliabadi M, Atique S, Ausloos M, Awasthi A, Quintanilla BPA, Azari S, Badawi A, Banoub JAM, Barker-Collo SL, Barnett A, Bedi N, Bennett DA, Bhattacharjee NV, Bhattacharyya K, Bhattarai S, Bhutta ZA, Bijani A, Bikbov B, Britton G, Burstein R, Butt ZA, Cárdenas R, Carvalho F, Castañeda-Orjuela CA, Castro F, Cerin E, Chang JC, Collison ML, Cooper C, Cork MA, Daoud F, Das Gupta R, Weaver ND, De Neve JW, Deribe K, Desalegn BB, Deshpande A, Desta M, Dhimal M, Diaz D, Dinberu MT, Djalalinia S, Dubey M, Dubljanin E, Durães AR, Dwyer-Lindgren L, Earl L, Kalan ME, El-Khatib Z, Eshrati B, Faramarzi M, Fareed M, Faro A, Fereshtehnejad SM, Fernandes E, Filip I, Fischer F, Fukumoto T, García JA, Gill PS, Gill TK, Gona PN, Gopalani SV, Grada A, Guo Y, Gupta R, Gupta V, Haj-Mirzaian A, Haj-Mirzaian A, Hamadeh RR, Hamidi S, Hasan M, Hassen HY, Hendrie D, Henok A, Henry NJ, Prado BH, Herteliu C, Hole MK, Hossain N, Hosseinzadeh M, Hu G, Ilesanmi OS, Irvani SSN, Islam SMS, Izadi N, Jakovljevic M, Jha RP, Ji JS, Jonas JB, Shushtari ZJ, Jozwiak JJ, Kanchan T, Kasaeian A, Karyani AK, Keiyoro PN, Kesavachandran CN, Khader YS, Khafaie MA, Khan EA, Khater MM, Kiadaliri AA, Kiirithio DN, Kim YJ, Kimokoti RW, Kinyoki DK, Kisa A, Kosen S, Koyanagi A, Krishan K, Defo BK, Kumar M, Kumar P, Lami FH, Lee PH, Levine AJ, Li S, Liao Y, Lim LL, Listl S, Lopez JCF, Majdan M, Majdzadeh R, Majeed A, Malekzadeh R, Mansournia MA, Martins-Melo FR, Masaka A, Massenburg BB, Mayala BK, Mehta KM, Mendoza W, Mensah GA, Meretoja TJ, Mestrovic T, Miller TR, Mini GK, Mirrakhimov EM, Moazen B, Mohammad DK, Darwesh AM, Mohammed S, Mohebi F, Mokdad AH, Monasta L, Moodley Y, Moosazadeh M, Moradi G, Moradi-Lakeh M, Moraga P, Morawska L, Morrison SD, Mosser JF, Mousavi SM,

Murray CJL, Mustafa G, Nahvijou A, Najafi F, Nangia V, Ndwandwe DE, Negoi I, Negoi RI, Ngunjiri JW, Nguyen CT, Nguyen LH, Ningrum DNA, Noubiap JJ, Shiadeh MN, Nyasulu PS, Ogbo FA, Olagunju AT, Olusanya BO, Olusanya JO, Onwujekwe OE, Ortega-Altamirano DDV, Ortiz-Panozo E, Øverland S, P. A. M, Pana A, Panda-Jonas S, Pati S, Patton GC, Perico N, Pigott DM, Pirsaheb M, Postma MJ, Pourshams A, Prakash S, Puri P, Qorbani M, Radfar A, Rahim F, Rahimi-Movaghar V, Rahman MHU, Rajati F, Ranabhat CL, Rawaf DL, Rawaf S, Reiner RC, Remuzzi G, Renzaho AMN, Rezaei S, Rezapour A, Rios-González C, Roever L, Ronfani L, Roshandel G, Rostami A, Rubagotti E, Sadat N, Sadeghi E, Safari Y, Sagar R, Salam N, Salamati P, Salimi Y, Salimzadeh H, Samy AM, Sanabria J, Santric Milicevic MM, Sartorius B, Sathian B, Sawant AR, Schaeffer LE, Schipp MF, Schwebel DC, Senbeta AM, Sepanlou SG, Shaikh MA, Shams-Beyranvand M, Shamsizadeh M, Sharafi K, Sharma R, She J, Sheikh A, Shigematsu M, Siabani S, Silveira DGA, Singh JA, Sinha DN, Skirbekk V, Sligar A, Sobaih BH, Soofi M, Soriano JB, Soyiri IN, Sreeramareddy CT, Sudaryanto A, Babale Sufiyan M, Sutradhar I, Sylaja PN, Tabarés-Seisdedos R, Tadesse BT, Temsah MH, Terkawi AS, Tessema B, Tessema ZT, Thankappan KR, Topor-Madry R, Tovani-Palone MR, Tran BX, Car LT, Ullah I, Uthman OA, Valdez PR, Veisani Y, Violante FS, Vlassov V, Vollmer S, Thu Vu G, Waheed Y, Wang YP, Wilkinson JC, Winkler AS, Local Burden of Disease Educational Attainment Collaborators (2020) Mapping disparities in education across low- and middle-income countries. *Nature* 577: 235–238. CrossRef

Hall RE (1978) Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence. *Journal of Political Economy* 86: 971–987. CrossRef

Hanson GH (1996) *U.S.-Mexico Integration and Regional Economies*. National Bureau of Economic Research, Cambridge. CrossRef

Jordahl K, Van Den Bossche J, Wasserman J, McBride J, Gerard J, Tratner J, Perry M, Farmer C, Cochran M, Gillies S, Bartos M, Culbertson L, Eubank N, Maxalbert, Fleischmann M, Hjelle GA, Arribas-Bel D, Ren C, Rey S, Journois M, Wolf LJ, Bilogur A, Grue N, Wilson J, YuichiNotoya, Wasser L, Filipe, Holdgraf C, Greenhall A, Trengrove J (2019) Geopandas/geopandas: V0.4.1. ZENODO software repository. CrossRef

Kanbur S, Venables A (2005) *Spatial Inequality and Development*. Oxford University Press, Oxford, UK. CrossRef

Khedmati Morasae E, Derbyshire DW, Amini P, Ebrahimi T (2024) Social determinants of spatial inequalities in COVID-19 outcomes across England: A multiscale geographically weighted regression analysis. *SSM - Population Health* 25: 101621. CrossRef

Knaap E (2017) The Cartography of Opportunity: Spatial Data Science for Equitable Urban Policy. *Housing Policy Debate* 27: 913–940. CrossRef

Ledić M, Rubil I, Urban I (2023) Tax progressivity and social welfare with a continuum of inequality views. *International Tax and Public Finance* 30: 1266–1296. CrossRef

Milanović B (2007) *Worlds Apart: Measuring International and Global Inequality*. Princeton University Press, Princeton (N.J.). CrossRef

Milanović B (2018) *Global Inequality: A New Approach for the Age of Globalization* (First Harvard University Press paperback ed.). The Belknap Press of Harvard University Press, Cambridge, Massachusetts London, England

Nijman J, Wei YD (2020) Urban inequalities in the 21st century economy. *Applied Geography* 117: 102188. CrossRef

OECD (2018) *Divided Cities: Understanding Intra-urban Inequalities*. OECD. CrossRef

Overman HG, Xu X (2022) Spatial disparities across labour markets. Institute for Fiscal Studies, London, UK. https://ifs.org.uk/books/spatial-disparities-across-labour-markets

Panzera D, Postiglione P (2020) Measuring the Spatial Dimension of Regional Inequality: An Approach Based on the Gini Correlation Measure. *Social Indicators Research* 148: 379–394. CrossRef

Partridge JS, Partridge MD, Rickman DS (1998) State patterns in family income inequality. *Contemporary Economic Policy* 16: 277–294. CrossRef

Pen J (1971) *Income distribution*. Allen Lane, London

Piketty T (2014) *Capital in the Twenty-first Century*. Harvard University Press. CrossRef

Piketty T, Saez E (2003) Income Inequality in the United States, 1913-1998. *The Quarterly Journal of Economics* 118: 1–39. CrossRef

Redding S, Venables AJ (2004) Economic geography and international inequality. *Journal of International Economics* 62: 53–82. CrossRef

Rey SJ (2015) Discrete regional distributional dynamics revisited. *Revue d'Économie Régionale & Urbaine* mai: 83–104. CrossRef

Rey SJ, Anselin L, Amaral P, Arribas-Bel D, Cortes RX, Gaboardi JD, Kang W, Knaap E, Li Z, Lumnitz S, Oshan TM, Shao H, Wolf LJ (2022) The PySAL Ecosystem: Philosophy and Implementation. *Geographical Analysis* 54: 467–487. CrossRef

Rey SJ, Arribas-Bel D, Wolf LJ (2023) *Geographic Data Science with Python*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, Boca Raton. CrossRef

Rey SJ, Casimiro Vieyra E (2023) Spatial inequality and place mobility in Mexico: 2000–2015. *Applied Geography* 152: 102871. CrossRef

Rey SJ, Montouri BD (1999) US regional income convergence: A spatial econometric perspective. *Regional Studies* 33: 143–156. CrossRef

Rey SJ, Sastré-Gutiérrez ML (2010) Interregional inequality dynamics in Mexico. *Spatial Economic Analysis* 5: 277–298. CrossRef

Rey SJ, Smith RJ (2013) A spatial decomposition of the Gini coefficient. *Letters in Spatial and Resource Sciences* 6: 55–70. CrossRef

Roemer JE (1998) *Equality of Opportunity*. Harvard Univ. Press, Cambridge, Mass. CrossRef

Saez E, Zucman G (2016) Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data *. *The Quarterly Journal of Economics* 131: 519–578. CrossRef

Sala-i-Martin X (2006) The World Distribution of Income: Falling Poverty and... Convergence, Period. *The Quarterly Journal of Economics* 121: 351–397. CrossRef

Sarkar S, Cottineau-Mugadza C, Wolf LJ (2024) Spatial inequalities and cities: A review. *Environment and Planning B: Urban Analytics and City Science* 51: 1391–1407. CrossRef

Sen A (2004) *Inequality Reexamined* (Reprint ed.). Oxford Univ. Press, New York. CrossRef

Suss J, Kemeny T, Connor DS (2024) GEOWEALTH-US: Spatial wealth inequality data for the United States, 1960–2020. *Scientific Data* 11: 253. CrossRef

Theil H (1967) *Economics and Information Theory*. North Holland, Amsterdam

Tine M (2017) Growing up in Rural vs. Urban Poverty: Contextual, Academic, and Cognitive Differences. In: *Poverty, Inequality and Policy*. IntechOpen. CrossRef

U.S. Department of Commerce (2023) Geographic Inequality on the Rise in the U.S. Regional economic research initiatve blog

Venter ZS, Figari H, Krange O, Gundersen V (2023) Environmental justice in a very green city: Spatial inequality in exposure to urban nature, air pollution and heat in Oslo, Norway. *Science of The Total Environment* 858: 160193. CrossRef

Wang J, Kwan MP, Xiu G, Deng F (2024) A robust method for evaluating the potentials of 15-minute cities: Implications for sustainable urban futures. *Geography and Sustainability*: S2666683924000646. CrossRef

Waskom M (2021) Seaborn: Statistical data visualization. *Journal of Open Source Software* 6: 3021. CrossRef

Wei R, Rey S, Knaap E (2020) Efficient regionalization for spatially explicit neighborhood delineation. *International Journal of Geographical Information Science*: 1–17. CrossRef

Widuto A (2019) Regional inequalities in the EU. European parliamentary research service, https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637951/EPRS_-BRI(2019)637951_EN.pdf

Zhang X, Kanbur R (2001) What difference do polarisation measures make?: An application to China. *Journal of Development Studies* 37: 85–98. CrossRef

REGION
The Journal of ERSA
Powered by WU

Volume 12, Number 1, 2025, 47–71
DOI: 10.18335/region.v12i1.516

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

Do local attitudes change with the exposure and the status of the migrants?

Bianca Biagi[1], Dionysia Lambiri[2], Marta Meleddu[1]

[1] University of Sassari and CRENoS (Italy), GSSI, L'Aquila, Italy
[2] Athens Coordination Centre for Migrant and Refugees, Athens, Greece

Received: 2 October 2023/Accepted: 2 March 2025

Abstract. Attitudes and perceptions regarding refugees and migrants play a vital role in the integration potential of newcomers and reflect policies and policy changes. This paper investigates how the exposure of urban communities to the presence of refugees and migrants in their local neighbourhoods affects their evaluation of the potential for migrant integration in the host country. Furthermore, it investigates the existence of a bias in the awareness of the presence of refugees and whether these evaluations change according to the status of the migrant. Using a unique dataset on the individual perceptions of residents of the Greek capital Athens, the analysis shows a positive effect of perceived presence and contends that perceptions of the size of refugee and migrant populations are more consequential for the formation of attitudes than the actual size. Moreover, residents tend to be more favourably disposed towards those recognised as refugees than they are towards permanent migrants.

JEL classification: F22, J60

Key words: migrants, refugees, exposure, integration, public opinions, perception bias

1 Introduction

In OECD countries, more than 5 million additional people migrate permanently (+ 7% in 2016 with respect to 2015; OECD 2018), and on average, more than 10% of residents are born abroad (Germany 15.7%; UK 13.4%, Greece 11.6%, Italy 10.4%). The 2030 Agenda for Sustainable Development of the United Nations recognises the importance of migration for sustainable development considering the "number of countries with migration policies to facilitate orderly, safe, regular, and responsible migration and mobility of people". Migration policies are set at a national level; however, it is the local context that matters when considering active measures for migrant integration as well as impacts on social policy, local labour markets, public services, and amenities.

The relationship between the presence of migrants and local economic performance is not straightforward; heterogeneity in cultural traits and level of education, the conditions under which net benefits prevail over costs, is still a research issue. Indeed, the level of integration strictly depends on the quantity/quality of migrants and natives as well as the perceived and actual cultural distance between them (Easterly, Levine 1997, Ottaviano, Peri 2006, Spies, Schmidt-Catran 2016, Bove, Elia 2017, Gradstein, Justman 2019).

The upsurge in anti-immigration sentiments has inflamed the policy debate throughout Europe (Bansak et al. 2016, Percoco, Fratesi 2018). Such public beliefs range broadly from generalised hostility towards immigration and a widespread fear over its perceived effects to scepticism around the possibility of integrating migrant populations in local communities while social cohesion is safeguarded. Understanding public attitudes towards migration and the underlying factors that drive them are central. Public attitudes determine policy changes (e.g., policy decisions on free-movement restrictions). They also influence collective visions and perceptions of who is considered a member of the in-group and who is not, affecting the potential for interaction as well as the prospects of conflict among different groups and, in turn, integration (Curtice 2017). Several factors shape public attitudes towards refugees and migrants, and the number of migrants is a crucial determinant (Bansak et al. 2016). However, studies have found that the perceived number of migrants overestimates the real numbers (Alesina et al. 2018, 2019, Steele, Perkins 2019). The misperception – when the size and the composition of migrants are seen differently than the actual numbers – might generate bias in public opinions. The intensity and direction of the relationship between misperception and attitudes towards migrants are not straightforward, and recent research offers mixed results. On the one hand, it supports the association between misperception and anti-immigrant attitudes (Pottie-Sherman, Wilkes 2017, Gorodzeisky, Semyonov 2019). On the other hand, it finds 1) a weak relationship between the objective and subjective evaluation of natives about the number of migrants and 2) a weak linkage between these subjective evaluations and attitudes towards integration (Spies, Schmidt-Catran 2016). Therefore, exposure to refugees and migrants in everyday life might positively or negatively affect perceptions. According to the intergroup theory proposed by Allport (1954), closer contact between natives and non-natives might reduce the prejudice towards minority groups and reduce extremism (Steinmayr 2020).

Interestingly, some studies have found that people tend to be more favourably disposed towards those recognised as refugees rather than other migrants (Mayda 2006, O'Rourke, Sinnott 2006, Hatton 2016). The word migration often implies a voluntary process, such as people who cross a border searching for better economic opportunities. This is not the case for refugees who cannot return to their homes in safe conditions and are consequently entitled to specific protection measures (UNHCR 2025).

Overall, previous research highlights the attitudes towards refugees and migrants and their subsequent integration, which is dependent on the socio-demographic and cultural characteristics of migrants, residents, the distance between them, and the contact between them.

This study starts with the premise that local attitudes and perceptions play a vital role in the integration potential of newcomers. Other fundamental structural factors are national integration policies that safeguard equal rights and access to services for migrant populations and local integration practices that aim to maximise opportunities for interaction. Specifically, this work concentrates on how the exposure of urban communities to the presence of refugees and migrant groups in their local neighbourhoods affects their evaluation of the potential for integration in the host country.

The first hypothesis (H1) is that the exposure to refugees (i.e., the possibility of interaction) reduces the negative attitudes and perceptions of the resident population towards them. Consequently, it might affect residents' beliefs about integration later on. The second hypothesis (H2) is that bias in the awareness of the presence of refugees may reinforce the residents' perception of the potential for integration. The third hypothesis (H3) is that the perceptions of integration may differ due to the status of refugees and migrants; while refugees are displaced due to conflict or persecution, migrants are free people who moved away from their country to seek better economic and educational opportunities.

In the present paper, the issue of integration focuses on the perspective of the resident population. The work uses a unique dataset on the individual perceptions of residents of the Greek capital Athens obtained less than two years after the outbreak of the refugee crisis in the summer of 2015. Between 2016 and 2017, the City of Athens Observatory for Refugees and Migrants (AORI) undertook a research programme consisting of a refugee

census and public opinion surveys to understand attitudes towards migrants and refugees. A challenge related to the situation is the increasing discontent among Greek nationals and existing migrant communities. As in the rest of Europe (Bansak et al. 2016), the mobilisation of funds and resources to manage the refugee crisis has fanned social tension (details on refugees' integration policies in Greece are provided in Skleparis 2018). The humanitarian response to the refugees' crisis affects the quality and breadth of social and welfare services for nationals. This work studies how the perceived presence of refugees affects residents' evaluation of integration potential and explores whether misperception occurs between the perceived presence and the actual number of refugees. Finally, it investigates whether the potential for integration changes according to the status of migrants compared to that of refugees.

The present paper contributes to the literature on the formation of public opinion of out-group populations in various ways. First, it provides evidence that exposure to refugees and migrants in local neighbourhoods positively affects individual attitudes related to immigration. This paper finds evidence that perceived presence has a more substantial effect on such attitudes than the actual presence of out-group populations and reports more positive attitudes towards newly arrived refugee populations than towards longer-term migrants living in the city. Such findings are extremely timely, as policies on immigration and refugees are often motivated by prevailing public attitudes. The outcomes of the present work can inform policy-relevant research that examines the complex bidirectional relationship between societal perceptions related to migration and current anti-immigration narratives.

The paper is structured as follows. Section 2 reviews the existing literature on public attitudes towards migration. Section 3 presents a case study of Athens, a city that found itself at the forefront of an unprecedented refugee crisis at the European level. Section 4 explains the unique dataset used in study (4.1) and presents the empirical model and the methodology (4.2). Section 5 illustrates the main results, with a specific analysis on the effect of perceived versus actual presence (5.1) and on the effect of economic migrants (5.2), providing various robustness checks (5.3). The last section presents conclusions, discusses limitations, and compiles the policy implications of this work.

2 Perceptions on migration

Research on public opinion regarding immigration has grown in recent decades due mainly to the rapid increase in the phenomenon. Hainmueller, Hopkins (2014) classified the literature on immigration opinions into two main strands: political economy and sociopsychological. The former analyses the impact of immigration on individuals according to labour market competition (Hainmueller, Hopkins 2015, Valentino et al. 2019, Chletsos, Roupakias 2019), welfare (Facchini, Mayda 2009, Schmidt-Catran, Spies 2019), and fiscal burden (Campbell et al. 2006, Dustmann, Preston 2007). The economic strand highlights several factors that can affect negative and positive perceptions of migrants held by native-born individuals related to both their macro-contexts (e.g., mixed schools, the employment rate of the region), and their social characteristics (e.g., the personal knowledge of migrants, the level of difficulty in paying bills, and the inaccurate perception of the actual numbers of migrants; Citrin et al. 1997, Eurobarometer 2018, OECD 2018). The so-called sociopsychological strand is rather heterogeneous and ranges from attitudes towards differences in race, religion, etc., to perceived threats to national identity, prejudice, and stereotypes and recognises the role of mass media on attitudes concerning immigration (Hainmueller, Hopkins 2014).

The attitudes and opinions of local communities regarding refugees and migrants depend on socio-cultural openness and play a key role in local integration policies. A strand of recent research focuses on the effects of residents' misperception on the opinion and attitude towards refugees and migrants (Pottie-Sherman, Wilkes 2017, Alesina et al. 2018, Steele, Perkins 2019, Gorodzeisky, Semyonov 2019). Overall, the findings confirm misperception and the linkage between misperception, anti-immigrant attitudes, and related policies (redistribution and welfare policies, and general social policies). In this context, Alesina et al. (2018, 2019) find that the perceived number of migrants is always

twice as high as reality for a set of countries (Germany, France, Italy, Sweden, the United Kingdom, and the United States). Steele, Perkins (2019), focusing on New York neighbourhoods, confirm overestimation, even at a lower intensity.

Exposure to migrants and refugees can positively or negatively affect the opinions of the resident population. Applied research finds negative opinions in cities and regions with low-and medium-income individuals, low-skilled natives working in the sector more exposed to migrants, non-college-educated individuals, women, right-wing voters, smaller, and less urban municipalities, municipalities with high unemployment, high immigrant shares, or past immigration settlements (Young et al. 2018, Palermo et al. 2022). Positive perceptions are found in cities and regions with younger individuals, high skills and college-educated individuals, left-wing voters, and more urban municipalities (Hainmueller, Hiscox 2007, Constant, Zimmermann 2009, Dahlberg et al. 2011, Alesina et al. 2018, Dustmann et al. 2019, OECD 2018).

Several studies show distinctions in public attitudes based on refugees' and migrants' characteristics. Evidence from the UK, for instance, suggests that people tend to default to negativity when asked about immigration, but are much less prone to do so when asked about specific groups of migrants (Ford 2011). In particular, people tend to be more favourably disposed towards those recognised as refugees than they are towards other migrants (Mayda 2006, O'Rourke, Sinnott 2006, Hatton 2016).

The present study investigates the integration potential of migrants and refugees from the perspective of the resident population. This work contributes to this line of research by analysing the presence of misperceptions and disentangling the different roles migrants and refugees might play in residents' opinions of integration potential. The case of Athens is the first study of Greece on this specific topic.

3 The city of Athens

Following the outbreak of the refugee crisis in the summer of 2015, Athens, the capital of Greece, found itself at the forefront of an unprecedented emergency at the European level. On top of Greece's domestic economic crisis, the influx of large numbers of refugees – mainly from Syria, Afghanistan, and Iraq – found the country unprepared to deal with complex challenges, which ranged from the provision of short-term accommodation solutions for asylum seekers to longer-term support for the efficient integration of recognised refugees and migrants into Greek society. United Nations High Commissioner for Refugees (UNHCR) data for Greece indicate that, as of October 2018, 58% (over 12,000) of refugees living in UNHCR's 'ESTIA' accommodation programme were living in Athens and the region of Attica (Papatzani 2020). An additional 6,323 people resided in six open reception facilities (open campsites), with one, the site of Eleonas, located very close to the city centre (UNHCR 2018).

Significant immigration flows are not a new phenomenon in Greece. Indeed, starting in the early 1990s and especially following the collapse of the communist regime, Greece received major waves of migrants from the Balkans, Central, and Eastern Europe, and the former Soviet Union. During the last decade, particularly since the beginning of the economic crisis in 2008, Greece has become a transit point and destination for migrants and asylum seekers arriving from Southeast Asia, Africa, and the Middle East.

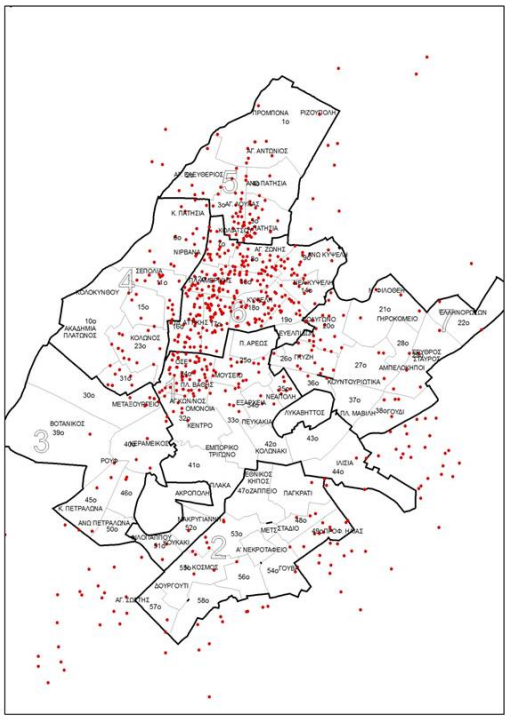
The largest nationality among migrants in Athens in 2016 was Albanians (38,469), followed in much smaller numbers by nationals from the Philippines, Bangladesh, and Ukraine (Table 1). There is no reliable information on the number of irregular migrants living in Athens. In terms of age, the majority of migrants in Athens are between 25 and 50 years old. In comparison, there is a significant age cohort among the younger generations between 0 and 14 years old – children born in Greece – that remain foreign nationals – or those who came to the country at a very early age.

The number of refugees and asylum seekers in Athens during 2016–2017 was estimated at 15,000 people (a share of over 40% of Greece's total number of refugees). It is worth noticing that, according to the 2011 census, migrants represent 17.7% of the total population in the Central Sector of the Prefecture of Attica (ELSTAT 2011).

According to preliminary observations, the district of Western Athens seems partic-

Table 1: Number of residence permits issued to third-country nationals in the Municipality of Athens, March 2016

Nationality	# of Permits	Nationality	# of Permits	Nationality	# of Permits
Albania	38,469	Moldova	2,120	Sri Lanka	499
Philippines	6,083	Syria	2,025	Ghana	475
Bangladesh	4,383	China	1,662	Armenia	452
Ukraine	4,026	Nigeria	1,194	Morocco	324
Egypt	3,549	Russia	1,186	Iran	312
Georgia	3,203	India	792	Other	3,258
Pakistan	3,068	Ethiopia	726	Total	77,806



Source: Public Issue, 2016.

Figure 1: Spatial distribution of refugee apartments in Athens

ularly concerned about migration with more than ten asylum seekers and refugees for every 1,000 people, compared to the average of more than four for every 1,000 people in the rest of Athens. On 30 April 2017, there were 98,107 recorded and pending asylum applications in Greece. Since then, asylum procedures have accelerated, but still challenge the public system, and a sizeable backlog remains (Proietti, Veneri 2021).

In Athens, as well as in other Greek cities, accommodation for asylum seekers and refugees is scarce. In the centre of Athens, once-abandoned urban spaces – mainly derelict retail spaces in the centre of the city – have been transformed into community centres offering services from language courses to legal representation and psychological support. By 2018, the Office of the United Nations High Commissioner for Refugees (UNHCR) had a housing programme for refugees – the ESTIA programme – and many refugees have found informal jobs and are renting apartments across the city, especially in multicultural neighbourhoods. Figure 1 shows the distribution of UNHCR accommodation apartments in the districts of Athens. The most significant concentration is in District 6 due to real estate availability under the UNHCR scheme. District 3 (Eleonas) hosts a temporary accommodation site.

Table 2: Distribution of interviews within Athens' city districts

District	Share	District	Share	District	Share
District 1	11%	District 4	13%	District 7	19%
District 2	16%	District 5	15%		
District 3	7%	District 6	20%		

Table 3: Sample description

Integration potential	
Cannot be integrated	40%
It depends	34%
Can be integrated	22%
Gender	
Male	53%
Female	47%
Age	
18–24	2%
25–34	4%
35–44	9%
45–54	18%
55–64	25%
>65	42%
Native	97%
Civil status	
Married with children under 18	14%
Married with children over 18	49%

Employment	
Employers/self-employed	9%
Public sector salaried employees	6%
Private sector salaried employees	13%
Unemployed	11%
Pensioners	49%
Housewives	9%
Students	1%
Other/no answer	2%
Financial situation	
Facing great difficulties	41%
Facing difficulties	34%
Making ends meet	22%
Living comfortably	3%
Political self-placement	
Left	17%
Centre	39%
Right	12%
Apolitical	32%

4 Methodological approach and empirical model

4.1 The data

Between 2016 and 2017, the AORI undertook a research programme consisting of a refugee census and public opinion surveys. Specifically, a public opinion survey aimed to understand the attitudes towards refugees and migrants of permanent residents of the city of Athens. The central questions concern the perceived presence, attitude towards coexistence, and integration of refugees. In 2016, a total of 3,024 residents aged 18 and over were interviewed in three waves of telephone surveys (1,007 in October, 1,012 in November, and 1,005 in December) by 22 interviewers and two supervisors. The sample was stratified according to the resident's neighbourhood. The standard error of the final sample is between +/- 3.2%, and the confidence interval was 95% (Table 2).

The question under analysis asks respondents to indicate their opinion about the integration potential of refugees: "Generally speaking, the refugees that remain in Greece, do you think that they can or they cannot be integrated into the Greek Society?" The dependent variable is a discrete variable that considers the respondent's perception of the possibility of refugees' integration. The response options are on a three-point Likert scale: 1 = cannot be integrated, 2 = it depends, and 3 = can be integrated. The majority of residents (40%) believe that refugees cannot be integrated, 22% believe that they can be integrated, and the remaining residents do not have a clear position. The majority of respondents are native, male, aged over 45, married with children over 18, pensioners, facing financial difficulties, and politically place themselves in the centre or left wings (Table 3).

4.2 The empirical model

As the dependent variable has more than two categories, and the values of each category have an expressive sequential order corresponding to the level of integration, the empirical analysis uses an ordered logit model. This model, also called the *proportional regressions*

model, implies that the observed ordinal variable Y is a function of a continuous latent variable, Y^*, which is not measured. Y^* has various threshold points, and the value of Y depends on whether a particular threshold is crossed (Menard 2002). Specifically, Y^* is equal to:

$$Y^* = \sum_{k=1}^{K} \beta_k X_{ki} + \varepsilon_i = Z_i + \varepsilon_i \tag{1}$$

where $Z_i = E(Y^*)$, and ε_i is the random disturbance term. Using the estimated value of Z and assuming a logistic distribution for the disturbance term, the ordered logit model estimates the probability that the unobserved variable Y^* falls within the various threshold limits. Furthermore, this specification assumes that the coefficients that express the relationship between the lowest threshold and all higher thresholds of the dependent variable are the same as those that describe the relationship between the next lowest category and all higher categories, and so on. In other words, because it is assumed that the relationship between all pairs of groups is the same, a single set of coefficients is estimated, and the parallel regression assumption holds. The empirical model applied in the present paper is as follows:

$$\begin{aligned} &\textit{Perception of refugees' integration}_i = f(\textit{Refugees' perceived presence}_i, \\ &\textit{Refugees' actual presence}_i, \textit{Immigrants' perceived presence}_i, \textit{Other controls}_i) \end{aligned} \tag{2}$$

Controls included individual socio-economic and demographic characteristics, such as gender, age, education, civil status, presence of children, employment, income adequacy (financial situation), and political self-placement. Furthermore, the controls included two variables that check for the perception that refugees might cause problems and that residents cannot distinguish between migrants and refugees. Variable descriptions are presented in Table A.1 in the Appendix.

The final model (Base Model) included the variables selected using a stepwise procedure. In this specification, the approximate likelihood-ratio test of proportionality of odds across response categories does not provide evidence that parallel regression assumption has been violated (chi2(16) = 15.04 and Prob > chi2 = 0.5219). This result is also confirmed by the Brant Test of Parallel Regression Assumption (chi2(16) = 22.40 and Prob > chi2 = 0.131). Therefore, the results can be interpreted by looking at the sign and significance of the coefficients.

5 Results

The present paper investigates three main hypotheses. First, that the exposure to refugees reduces the resident population's negative attitudes towards and perceptions of them (H1). Second, that a bias in the awareness of the presence of refugees may reinforce the perception of the potential for integration (H2). Third, that the perception of integration may differ due to the status of refugees and migrants (H3). Table 4 shows that *refugees' perceived presence* and *refugees' actual presence* are positive and significant (Model 1 and Model 2). This finding corroborates H1; hence, exposure (perceived and actual) to refugees reduces the resident population's negative attitudes towards them. The comparison between the coefficients of the variables *refugees' perceived presence* and *refugees' actual presence* confirms H2, as the effect of perception is stronger than the actual presence. This could be interpreted as a sign of misperception, confirming that perceptions are often stronger than actual facts (Alesina et al. 2018, Steele, Perkins 2019). Furthermore, when the *perception that foreigners cause problems* increases, opinions of integration potential decrease accordingly. Perception of the presence of migrants (*migrants' perceived presence)* negatively affects individual evaluations of the potential for integration. In line with H3, this result suggests that the status of refugees and migrants might affect integration perceptions. This might also suggest that refugees are perceived differently than migrants. According to previous research, residents tend to be more

Table 4: Residents' perception of refugees' integration potential

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
Refugees' perceived presence	0.103**	
	(0.0517)	
Refugees' actual presence		0.000434**
		(0.000212)
Perception that foreigners cause problems	-0.564***	-0.539***
	(0.0617)	(0.0557)
Migrants' perceived presence	-0.00335	-0.00314*
	(0.00208)	(0.00190)
Unable to distinguish between migrants/refugees	-0.221	-0.156
	(0.166)	(0.139)
Gender	-0.0491	-0.0750
	(0.0915)	(0.0862)
Age	-0.724***	-0.672***
	(0.208)	(0.198)
Age^2	0.0766***	0.0689***
	(0.0264)	(0.0251)
Education	0.208***	0.176**
	(0.0759)	(0.0708)
Married with children over 18	-0.0700	-0.0212
	(0.105)	(0.0979)
Married with children under 18	-0.142	-0.135
	(0.146)	(0.140)
Unemployed	0.0385	-0.0277
	(0.153)	(0.144)
Inactive	-0.133	-0.204
	(0.138)	(0.131)
Income adequacy	0.0940*	0.117**
	(0.0537)	(0.0506)
Born in Greece	-0.477*	-0.406
	(0.277)	(0.261)
Political self-placement (left)	1.034***	1.049***
	(0.123)	(0.117)
Political self-placement (centre)	0.172*	0.193**
	(0.101)	(0.095)
N	2164	2433
Pseudo R^2	0.071	0.071
AIC	3863.5	4356.8
BIC	3965.7	4461.1

Standard errors are in parentheses; * p <0.10, ** p <0.05, *** p <0.01.

favourably disposed towards individuals recognised as refugees than they are towards migrants (Mayda 2006, O'Rourke, Sinnott 2006, Hatton 2016).

Among the socio-economic and demographic controls, residents' age negatively affects perceptions of integration. The effect is not linear, indicating that younger residents have positive opinions (Age^2). Age, education level, and financial difficulties affect public opinion (Card et al. 2005, Mayda 2006, O'Rourke, Sinnott 2006, Hainmueller, Hiscox 2007, 2010, Alesina et al. 2018, 2019, Hatton 2020). Residents *born in Greece* are found to be more sceptical about refugees' integration potential than non-natives (see Model 1). Finally, confirming previous findings, political self-placement affects integration perceptions. Specifically, residents who vote for left-wing and centre political parties have a favourable opinion about integration (Dustmann et al. 2019, Alesina et al. 2018, 2019). Tables A.4 and A.5 in the Appendix present the marginal effects for both models across each threshold of the dependent variable (i.e., *Cannot be integrated, Depends and Can be integrated*).

5.1 The effects of perceived versus actual presence

Existing literature suggests that perceptions often play a bigger role than facts in how views are formed. Specifically, Alesina et al. (2018) found that the perceived number of

Table 5: The effect of refugees' perceived presence versus actual presence: marginal effects expressed in percentages

Dependent: Perception of refugees' integration	Cannot be integrated pr(y = 1)	Depends	Can be integrated pr(y = 3)
Refugees' perceived presence	-2.230%**	0.313%**	1.920%**
	(0.0111)	(0.00158)	(0.00959)
Refugees' actual presence	-0.009%**	0.001%**	0.008%**
	(0.0000457)	(0.00000668)	(0.0000391)

Notes: Standard errors are in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: The effect of refugees' perceived presence: marginal effects expressed in percentage by categories

Dependent: Perception of refugees' integration Refugees' perceived presence	Cannot be integrated pr(y = 1)	Depends	Can be integrated pr(y = 3)
None	58.70%	13.30%	28.00%
A few	56.50%	13.60%	29.90%
Some	54.30%	13.90%	31.80%
Many	52.00%	14.10%	33.80%

migrants is always two times higher than reality. A more in-depth analysis on this issue in the case of Athens compares the marginal effects of the *refugees' perceived presence* and *refugees' actual presence* variables (Table 5). When transforming the coefficients into percentages, the present analysis confirms the impact of perception over actual presence, as in Steele, Perkins (2019) and, specifically, a perception of double the number present in reality, as in Alesina et al. (2018).

Investigating in more detail how perceived presence affects the perception of integration potential, Figures 2 and 3 show refugees' role in the neighbourhood. Figure 2 (Table 6) compares *refugees' perceived presence* with the *perception of refugees' integration*; the dashed line shows that the predicted probability of the perception of integration (the y-axis) goes from 28% – when the residents are not at all exposed to refugees in their neighbourhood – to 34% – the maximum exposure. The line continuously moves in the same direction: the probability of no integration decreases as exposure increases (the predicted probability goes from 59% to 52%).

The same results were confirmed when analysing the effect of the actual presence on predicted probabilities (Figure 3, Table 7). Overall, the findings indicate that the higher the opportunities to interact with refugees, the higher the residents' positive opinion on refugees' integration.

Table 7: The effect of refugees' actual presence: marginal effects expressed in percentage by categories

Dependent: Perception of refugees' integration Refugees' actual presence	Cannot be integrated pr(y = 1)	Depends	Can be integrated pr(y = 3)
2%	58.60%	13.80%	27.70%
8%	57.80%	13.90%	28.30%
9%	57.80%	13.90%	28.40%
11%	57.50%	13.90%	28.60%
13%	57.20%	14.00%	28.80%
50%	53.00%	14.50%	32.50%

Notes: Refugees' actual presence is the percentage of total refugees hosted in each city district.

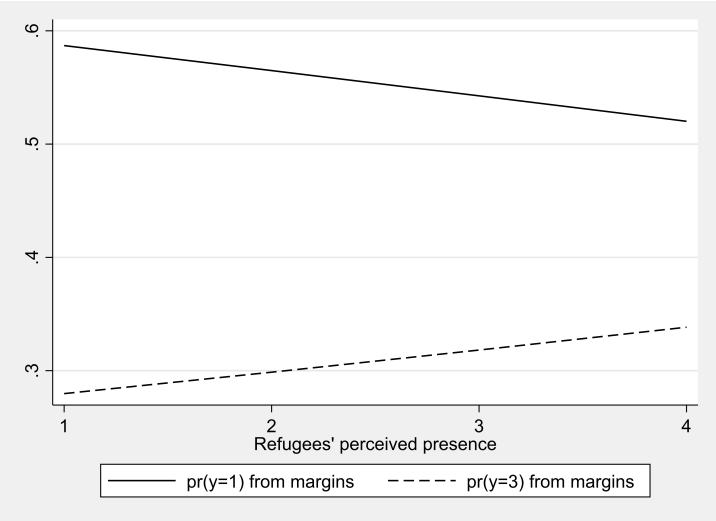


Figure 2: The effect of refugees' perceived presence

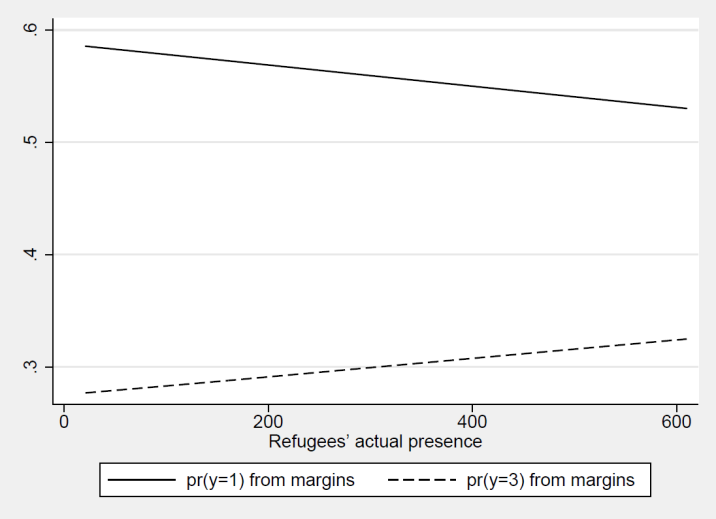


Figure 3: The effect of refugees' actual presence

5.2 The effect of migrants' perceived presence

Another result regards the role of the perceived presence of migrants on residents' opinions. The migrants' impact is not positive (Figure 4, Table 8). Indeed, the predicted probability of integration decreases as the perceived presence of migrants increases – the dashed line in Figure 4 shows that the probability goes from 32% to 26% – while the probability of no integration increases as the migrants' perceived presence increases – the solid line in Figure 4 shows that the probability goes from 54% to 61%.

This result might also indicate that migrants are not fully integrated into Athens. Therefore, their perceived presence in each district might negatively affect residents opinions on the prospective integration of refugees. Furthermore, residents would likely perceive refugees as more educated than migrants and, therefore, more likely to be integrated into the local context. Indeed, previous literature has found that cultural adaptability relates to the level of education (Algan et al. 2012). Unfortunately, no information about refugees' education levels is available. Moreover, this result could also be capturing one of the first effects of the ad hoc integration policy implemented in Athens after the first refugee crisis in 2015 (Skleparis 2018).

Table 8: The effect of migrants' perceived presence: marginal effects expressed in percentage by categories

Dependent: Perception of refugees' integration Migrants' perceived presence	Cannot be integrated pr(y = 1)	Depends	Can be integrated pr(y = 3)
1 %	54.30%	14.10%	31.60%
5 %	54.60%	14.00%	31.30%
8 %	54.80%	14.00%	31.10%
15 %	55.40%	14.00%	30.70%
30 %	56.50%	13.80%	29.70%
50 %	57.90%	13.60%	28.50%
80 %	60.10%	13.20%	26.70%
90 %	60.80%	13.10%	26.10%
98 %	61.40%	13.00%	25.60%

Notes: Migrants' perceived presence is the proportion of foreigners living in the city district as a subjective estimation.

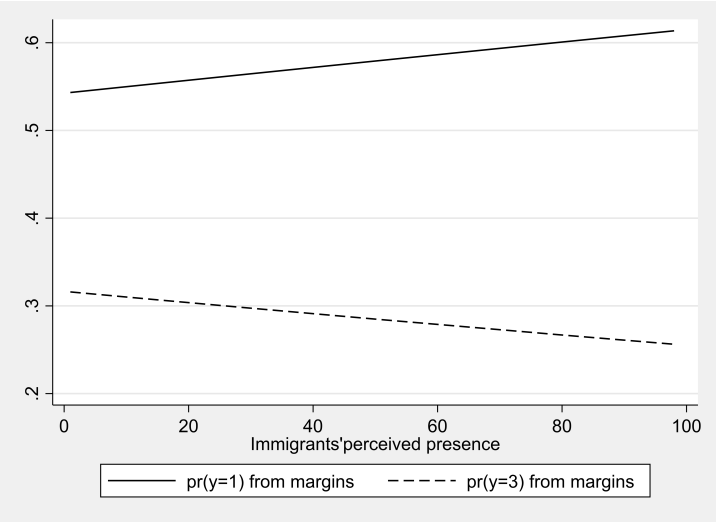


Figure 4: The effect of migrants' perceived presence

Table 9: Robustness check. Including the new migrants' perceived presence and new perception that foreigners cause problems

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
Refugees' perceived presence	0.0940*	
	(0.0498)	
Refugees' actual presence		0.000384*
		(0.000199)
Perception that foreigners cause problems 1		
Perception that foreigners cause problems 2	-0.554***	-0.537***
	(0.105)	(0.0975)
Perception that foreigners cause problems 3	-1.271***	-1.182***
	(0.157)	(0.141)
Perception that foreigners cause problems 4	-1.632***	-1.550***
	(0.249)	(0.224)
New Migrants' perceived presence	-0.109	-0.110
	(0.138)	(0.121)
Unable to distinguish between	-0.193	-0.130
migrants/refugees	(0.162)	(0.133)
Gender	-0.0282	-0.0434
	(0.0897)	(0.0836)

continued on next page ...

Table 9: Robustness check. Including the new migrants' perceived presence and new perception that foreigners cause problems – continued

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
Age	-0.725***	-0.652***
	(0.214)	(0.202)
Age^2	0.0771***	0.0667***
	(0.0272)	(0.0256)
Education	0.212***	0.175***
	(0.0725)	(0.0675)
Married with children over 18	-0.0865	-0.0459
	(0.101)	(0.0934)
Married with children under 18	-0.140	-0.123
	(0.147)	(0.140)
Unemployed	0.0781	-0.00330
	(0.151)	(0.141)
Inactive	-0.101	-0.168
	(0.144)	(0.133)
Income adequacy	0.0802	0.102**
	(0.0529)	(0.0493)
Born in Greece	-0.363	-0.382
	(0.268)	(0.247)
Political self-placement (left)	1.066***	1.069***
	(0.120)	(0.113)
Political self-placement (centre)	0.194*	0.200**
	(0.0998)	(0.0928)
Cut1	-1.075**	-1.163**
	(0.510)	(0.474)
Cut2	-0.396	-0.457
	(0.509)	(0.473)
N	2236	2537
Pseudo R^2	0.070	0.068
AIC	4004.3	4566.1
BIC	4118.5	4682.9

Notes: Robust standard errors are in parentheses; *p < 0.10, **p < 0.05, ***p < 0.01.

5.3 Robustness check

Some variables related to the perception of refugees and migrants are weakly correlated. The correlation table in the appendix (Table A.2) shows that the most correlated variables are: *refugees' perceived presence* and *migrants' perceived presence*; *migrants' perceived presence* and *perception that foreigners cause problems*. Therefore, to check whether the results hold, migrants*' perceived presence* and *perception that foreigners cause problems* were transformed into dummy variables. In particular, the continuous variable *migrants' perceived presence* has been transformed into a dummy variable that takes the value 1 if the share of foreigners over the total residents in the district is higher than 75%, and 0 otherwise. Other dummy variables with different thresholds have been tried (> 25; > 55; > 70) and the least correlated one resulted in the > 75 threshold. The correlation of n*ew migrants' perceived presence* and *refugees' perceived presence* reduces to 0.19 (originally it was 0.39, compare Table A.2 and Table A.3 in the Appendix). Furthermore, the *perception that foreigners cause problems* has been split into 4 dummies that take the following values depending on the response options: 1 = none, 2 = a few, 3 = some, 4 = many, and 0 otherwise. This transformation reduces the correlation between the *perception that foreigners cause problems* and n*ew migrants' perceived presence* (compare Tables A.2 and A.3 in the Appendix). As a further check, we also estimated two additional models, transforming all categorical variables of Model 1 and Model 2 into dummy variables. As shown in Table A.6, the results align with previous findings. Table 9 shows that the results also remained stable using the two transformed variables. A set of regressions controls for the fixed effects of ethnic nationalities and residents' neighbourhood location. Table 10 shows that refugees' perceived presence and refugees' actual presence remain consistently stable.

Table 10: Robustness check by nationality of refugees and migrants

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
Albanians		
Refugees' perceived presence	0.111**	
	(0.0526)	
Refugees' actual presence		0.000446**
		(0.000212)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3864.8	4358.3
BIC	3972.7	4468.4
Pakistanis		
Refugees' perceived presence	0.109**	
	(0.0520)	
Refugees' actual presence		0.000418**
		(0.000212)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3864.1	4357.1
BIC	3972.0	4467.2
Africans		
Refugees' perceived presence	0.101*	
	(0.0518)	
Refugees' actual presence		0.000361*
		(0.000218)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3862.8	4356.9
BIC	3970.8	4467.0
Filipinos		
Refugees' perceived presence	0.0938*	
	(0.0524)	
Refugees' actual presence		0.000406*
		(0.000214)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3864.3	4357.9
BIC	3972.2	4468.1
Syrians		
Refugees' perceived presence	0.0968*	
	(0.0538)	
Refugees' actual presence		0.000429**
		(0.000212)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.071	0.071
AIC	3865.3	4358.3
BIC	3973.2	4468.4

Notes: Standard errors are in parentheses; *p < 0.10, **p < 0.05, ***p < 0.01.

As explained in Section 3, the city of Athens is divided into seven districts. Results stay stable for all districts; the only exception are districts six and seven, where actual presence does not affect the residents' opinion of integration potential in the neighbourhood (Table 11).

Table 11: Robustness check by neighbourhood (districts)

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
District 1		
Refugees' perceived presence	0.104**	
	(0.0519)	
Refugees' actual presence		0.000446**
		(0.000212)
District 1	-0.0204	0.0678
	(0.149)	(0.141)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3864.8	4358.3
BIC	3972.7	4468.4
District 2		
Refugees' perceived presence	0.104**	
	(0.0519)	
Refugees' actual presence		0.000443**
		(0.000217)
District 2	0.0165	0.0235
	(0.120)	(0.116)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3864.1	4357.1
BIC	3972.0	4467.2
District 3		
Refugees' perceived presence	0.103**	
	(0.0517)	
Refugees' actual presence		0.000472**
		(0.000218)
District 3	0.0279	0.118
	(0.172)	(0.165)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3862.8	4356.9
BIC	3970.8	4467.0
District 4		
Refugees' perceived presence	0.102**	
	(0.0518)	
Refugees' actual presence		0.000520**
		(0.000219)
District 4	0.222	0.212
	(0.138)	(0.131)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3864.3	4357.9
BIC	3972.2	4468.1
District 5		
Refugees' perceived presence	0.104**	
	(0.0517)	
Refugees' actual presence		0.000419**
		(0.000214)
District 5	-0.150	-0.0636
	(0.127)	(0.121)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.071	0.071
AIC	3865.3	4358.3
BIC	3973.2	4468.4
District 6		
Refugees' perceived presence	0.0943*	

continued on next page ...

Table 11: Robustness check by neighbourhood (districts) – continued

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
	(0.0521)	
Refugees' actual presence		-0.000262
		(0.00137)
District 6	0.172	0.358
	(0.115)	(0.695)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3862.8	4356.9
BIC	3970.8	4467.0
District 7		
Refugees' perceived presence	0.0909*	
	(0.0522)	
Refugees' actual presence		0.000335
		(0.000218)
District 7	-0.206*	-0.214*
	(0.113)	(0.110)
Other controls	YES	YES
N	2164	2433
Pseudo R^2	0.072	0.071
AIC	3864.3	4357.9
BIC	3972.2	4468.1

As a final check, we address the potential joint endogeneity of the perception variables by estimating a reduced-form equation with only exogenous variables as regressors. Specifically, we use a binary logit model, where the dependent variable is *Can be integrated* (coded as 1 for "Can be integrated" and 0 otherwise). The independent variables include only strictly exogenous individual characteristics, omitting perception/opinion variables and focusing on *refugees' actual presence* as the main variable of interest. As shown in Table 12, the presence of refugees increases the likelihood that residents report that refugees can be integrated.

6 Conclusions and policy implications

This study is based on the premise that – in practice – integration takes place at the local level, as cities are focal locations for the refugee and migrant reception and integration processes. Additionally, although migration policies are the responsibility of national governments, the concentration of migrants in cities and metropolitan areas more broadly has a significant impact on local demands for labour, housing, and goods and services, creating challenges that fall to local authorities to manage (Boulant et al. 2016, Diaz Ramirez et al. 2018). The present paper analyses how urban communities' exposure to refugee and migrant groups in their local neighbourhoods affects their evaluation of the refugees' potential for integration into the host communities. Specifically, it explores how the exposure to refugees affects residents' evaluation of integration potential, whether misperception occurs between the perceived presence and the actual number, and to what extent the potential for integration changes according to migrant versus refugee status. Overall, the results corroborate the few existing studies on the positive effect of exposure (Steele, Perkins 2019) and contend that perceptions of the size of refugee and migrant populations are more consequential to the formation of attitudes related to refugees and migrants than is the actual size (Alesina et al. 2018, Gorodzeisky, Semyonov 2019). Moreover, in accordance with previous research, residents tend to be more favourably disposed towards refugees than they are towards permanent migrants (Mayda 2006, O'Rourke, Sinnott 2006, Hatton 2016).

Immigration policy-making is often motivated by prevailing public attitudes. Simultaneously, public opinion can be shaped by the ways in which political actors frame the issues and challenges at hand. Understanding public attitudes in host communities is an increasingly important task. One of the most crucial policy implications relates to

Table 12: Binary Logit Model: Reduced-Form Analysis

	Can be integrated
Refugees' actual presence	0.000146**
	(0.0000680)
Gender	0.0697
	(0.0481)
Age	-0.212***
	(0.0363)
Education	0.253***
	(0.0436)
Married with children over18	0.0103
	(0.0897)
Married with children under 18	-0.234*
	(0.134)
cons	-0.580***
	(0.164)
N	2856
pseudo R^2	0.018
AIC	3346.6
BIC	3382.3

the powers of perception and public opinion, which are as important as planning for an inclusive city. However, ensuring that public spaces are designed and utilised for meaningful encounters is critical. Proximity in neighbourhoods is insufficient to bring about positive inter-group attitudes without targeted work to bring different people together (Ahmed 2000). Social projects that allow locals and migrants to come together enable sustained and meaningful interactions, which more effectively generate positive intergroup attitudes (Matejskova, Leitner 2011) towards cultural diversity and spill over onto economic outcomes.

Several limitations of this study need to be acknowledged. These are mainly related to the nuances of the term 'integration', as interviewees can interpret it in various ways. More attention to public opinions and perceptions is needed from local and national policy advocates in Greece. Additional empirical research is required to understand the social dynamics that shape the subjective dimensions of the social integration of migrants and refugees.

As this work mainly relies on survey-based data, it does not capture the nuanced experiences of residents, which would have provided a deeper understanding of how perceptions are formed. Additionally, it is important to note that the sample overrepresents individuals aged 45 and above, which may introduce potential bias. However, this may reflect the demographic profile of the population residing in the neighbourhoods, as the sample is stratified by district. Furthermore, future research should also consider the cultural aspects and its barriers in order to better understand the mechanisms underlying integration issues. Future research should be complemented by a qualitative approach to allow for a more accurate interpretation of the socio-cultural determinants of perceptions and the narratives that shape them for both local and migrant residents.

References

Ahmed S (2000) *Strange Encounters: Embodied Others in Post-Coloniality*. Routledge, New York. CrossRef

Alesina A, Miano A, Stantcheva S (2018) Immigration and redistribution. NBER working paper, 24733. CrossRef

Alesina A, Murard E, Rapoport H (2019) Immigration and preferences for redistribution in Europe. NBER working paper, 25562. CrossRef

Algan Y, Bisin A, Manning A, Verdier T (2012) *Cultural Integration of Immigrants in Europe*. Oxford University Press, Oxford. CrossRef

Allport GW (1954) *The Nature of Prejudice*. Addison-Wesley, Cambridge, MA

Bansak K, Hainmueller J, Hangartner D (2016) How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science* 354: 217–222. CrossRef

Boulant J, Brezzi M, Veneri P (2016) *Income levels and inequality in metropolitan areas: A comparative approach in OECD countries*. OECD Publishing, Paris. CrossRef

Bove V, Elia L (2017) Migration, diversity, and economic growth. *World Development* 89: 227–239. CrossRef

Campbell AL, Citrin J, Wong C (2006) "Racial threat", partisan climate, and direct democracy: Contextual effects in three California initiatives. *Political Behavior* 28: 129–150. CrossRef

Card D, Dustmann C, Preston I (2005) Understanding attitudes to immigration: The migration and minority module of the first European social survey. CReAM DP, 0305

Chletsos M, Roupakias S (2019) Do immigrants compete with natives in the Greek labour market? Evidence from the skill-cell approach before and during the great recession. *The B.E. Journal of Economic Analysis & Policy* 19. CrossRef

Citrin J, Green D, Muste C, Wong C (1997) Public opinion toward immigration reform: The role of economic motivations. *The Journal of Politics* 59: 858–881. CrossRef

Constant AK, Zimmermann K (2009) Migration, ethnicity and economic integration. Technical report, IZA Discussion Papers, No. 4620. CrossRef

Curtice J (2017) Brexit: The vote to leave the EU. Litmus test or lightning rod? British social attitudes (34). NatCen Social Research, London

Dahlberg M, Edmark K, Lundqvist H (2011) Ethnic diversity and preferences for redistribution. CESIFO working paper, No. 3325. CrossRef

Diaz Ramirez M, Liebig T, Thoreau C, Veneri P (2018) The integration of migrants in OECD regions: A first assessment. OECD regional development working papers, No. 2018/01, OECD Publishing, Paris. CrossRef

Dustmann C, Preston I (2007) Racial and economic factors in attitudes to immigration. *The B.E. Journal of Economic Analysis & Policy* 7: 62. CrossRef

Dustmann C, Vasiljeva K, Damm AP (2019) Refugee migration and electoral outcomes. *Review of Economic Studies* 86: 2035–2091. CrossRef

Easterly W, Levine R (1997) Africa's growth tragedy: Policies and ethnic divisions. *The Quarterly Journal of Economics*: 1203–1250. CrossRef

ELSTAT (2011) Permanent population results. Table – 2011 Census. Available from: https://www.statistics.gr/en/2011-census-pop-hous

Eurobarometer (2018) Integration of immigrants in the European Union. https://population-europe.eu/books-and-reports/special-eurobarometer-469-integration-immigrants-european-union

Facchini G, Mayda AM (2009) Does the welfare state affect individual attitudes toward immigrants? Evidence across countries. *Review of Economics and Statistics* 91: 295–314. CrossRef

Ford R (2011) Acceptable and unacceptable immigrants: How opposition to immigration in Britain is affected by migrants' region of origin. *Journal of Ethnic and Migration studies* 37: 1017–1037. CrossRef

Gorodzeisky A, Semyonov M (2019) Perceptions and misperceptions: Actual size, perceived size and opposition to immigration in European societies. *Journal of Ethnical Migration Studies*: 1–19. CrossRef

Gradstein M, Justman M (2019) Cultural interaction and economic development: An overview. *European Journal of Political Economy* 59: 243–251. CrossRef

Hainmueller J, Hiscox MJ (2007) Educated preferences: Explaining attitudes toward immigration in Europe. *International Organization* 61: 399–442. CrossRef

Hainmueller J, Hiscox MJ (2010) Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment. *American Political Science Review* 104. CrossRef

Hainmueller J, Hopkins DJ (2014) Public attitudes toward immigration. *Annual Review of Political Science* 17: 225–49. CrossRef

Hainmueller J, Hopkins DJ (2015) The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. *American journal of political science* 59: 529–548. CrossRef

Hatton T (2020) Public opinion on immigration in Europe: Preference and salience. *European Journal of Political Economy*. CrossRef

Hatton TJ (2016) Immigration, public opinion and the recession in Europe. *Economic Policy* 86: 205–246. CrossRef

Matejskova T, Leitner H (2011) Urban encounters with difference: The contact hypothesis and immigrant integration projects in eastern Berlin. *Social & Cultural Geography* 12: 717–741. CrossRef

Mayda AM (2006) Who is against immigration? A cross-country investigation of individual attitudes toward immigrants. *Review of Economic Studies* 88: 510–530. CrossRef

Menard S (2002) *Longitudinal Research*. Quantitative Applications in the Social Sciences. SAGE Publications, Thousand Oaks, CA. CrossRef

OECD (2018) *Working Together for Local Integration of Migrants and Refugees*. OECD Publishing, Paris. CrossRef

O'Rourke K, Sinnott R (2006) The determinants of individual attitudes towards immigration. *European journal of political economy* 22: 838–861. CrossRef

Ottaviano G, Peri G (2006) The economic value of cultural diversity: Evidence from US cities. *Journal of Economic Geography* 6: 9–44. CrossRef

Palermo F, Sergi BS, Sironi E (2022) Does urbanization matter? Diverging attitudes toward migrants and Europe's decision-making. *Socio-Economic Planning Sciences* 101278. CrossRef

Papatzani E (2020) The geography of the 'ESTIA' accommodation program for asylum seekers in Athens. Athens social atlas, https://www.athenssocialatlas.gr/en/article/-the-geography-of-the-estia-accommodation-program/

Percoco M, Fratesi U (2018) The geography of asylum seekers and refugees in Europe (105–123). In: Biagi B, Faggian A, Rajbhandari I, Venhorst VA (eds), *New Frontiers in Interregional Migration Research*. Springer, Cham. CrossRef

Pottie-Sherman Y, Wilkes R (2017) Does size really matter? On the relationship between immigrant group size and anti-immigrant prejudice. *International Migration Review* 51: 218–250. CrossRef

Proietti P, Veneri P (2021) The location of hosted asylum seekers in OECD regions and cities. *Journal of Refugee Studies* 34: 1243–1268. CrossRef

Schmidt-Catran AW, Spies DC (2019) Immigration and welfare support in Germany: Methodological reevaluations and substantive conclusions. *American Sociological Review* 84: 764–768. CrossRef

Skleparis D (2018) Refugee integration in mainland Greece: Prospects and challenges. Policy brief, 02, Yasar University UNESCO chair on International Migration, https://-unescochair.yasar.edu.tr/wp-content/uploads/2018/02/Dimitris_PB02March2018.pdf

Spies D, Schmidt-Catran A (2016) Migration, migrant integration and support for social spending: The case of Switzerland. *Journal of European Social Policy* 26: 32–47. CrossRef

Steele l, Perkins K (2019) The effects of perceived neighborhood immigrant population size on preferences for redistribution in New York City: A pilot study. *Frontiers in sociology* 4: 18. CrossRef

Steinmayr A (2020) *Contact versus Exposure: Refugee Presence and Voting for the Far-Right*, Volume 103. Review of Economics and Statistics. CrossRef

UNHCR (2018) Inter-agency participatory assessment report - Greece 2018. United Nations High Comissioner for Refugees, inter-agency participatory assessment report, https://data.unhcr.org/en/documents/download/66441

UNHCR (2025) The 1951 Refugee Convention. Web page, https://www.unhcr.org/-about-unhcr/overview/1951-refugee-convention

Valentino NA, Soroka S, Iyengar S, Aalberg T, Duch R, Fraile M, Hahn K, Hansen KM, Harell A, Helbling M, Jackman SD, Kobayashi T (2019) Economic and cultural drivers of immigrant support worldwide. *British Journal of Political Science* 49: 1201–1226. CrossRef

Young Y, Loebachb P, Kim K (2018) Building walls or opening borders? Global immigration policy attitudes across economic, cultural and human security contexts. *Social Science Research* 75: 83–95. CrossRef

A Appendix:

Table A.1: Description of the variables

Variable name	Variable description	Source
Perception of refugees' integration	Discrete var. that takes into account the respondents' perception on the possibility of refugees' integration. The response options are: 1 = cannot be integrated, 2 = depends, and 3 = can be integrated.	AORI survey data
Refugees' perceived presence	Discrete var. that takes into account the perception of refugees' presence in the respondent's residential area. The response options are: 1 = none, 2 = a few, 3 = some, and 4 = many.	AORI survey data
Refugees' actual presence	Continuous var. that takes into account the number of refugees hosted in each city district.	Public Issue, 2016
Perception that foreigners cause problems	Discrete var. that takes into account the residents' perceptions about problems caused by foreigners in the residential area. The response options are: 1 = none, 2 = a few, 3 = some, and 4 = many.	AORI survey data
Migrants' perceived presence	Continuous var. that takes into account the proportion of foreigners living in the city district as a subjective estimation.	AORI survey data
Unable to distinguish between migrants/refugees	Dichotomous var. that takes a value of 1 if the respondent is unable to distinguish migrants from refuges; 0 otherwise	AORI survey data
Gender	Dichotomous var. that takes a value of 1 if male; 0 otherwise	AORI survey data
Age	Discrete var. that accounts for the respondent's age range. The response options are: 1 = 18–24, 2 = 25–34, 3 = 35–44, 4 = 45–54, 5 = 55–64, and 6 = >65.	AORI survey data
Age^2	The square of the respondent's age.	AORI survey data
Education	Discrete var. that takes into account the respondent's level of education. The response options are: 1 = primary, 2 = secondary, and 3 = tertiary.	AORI survey data
Married with children over 18	Dichotomous var. that takes a value of 1 if the respondent is married and has children over 18; 0 otherwise.	AORI survey data
Married with children under 18	Dichotomous var. that takes a value of 1 if the respondent is married and has children under 18; 0 otherwise.	AORI survey data
Unemployed	Dichotomous var. that takes a value of 1 if the respondent is unemployed at the time of the interview; 0 otherwise.	AORI survey data
Inactive	Dichotomous var. that takes a value of 1 if the respondent is inactive (i.e. pensioners, housewives, and students) at the time of the interview; 0 otherwise.	AORI survey data
Income adequacy	Discrete var. that takes into account the respondent's self-assessment of their personal financial situation. The response options are: 1 = facing great difficulties, 2 = facing difficulties, 3 = making ends meet, and 4 = living comfortably.	AORI survey data
Born in Greece	Dichotomous var. that takes a value of 1 if the respondent is a Greek native; 0 otherwise.	AORI survey data
Political self-placement (left)	Dichotomous var. that takes a value of 1 if the respondent declares that they belong to left-leaning political parties; 0 otherwise.	AORI survey data
Political self-placement (centre)	Dichotomous var. that takes a value of 1 if the respondent declares that they belong to centre political parties; 0 otherwise.	AORI survey data
District	Dichotomous var. that takes a value of 1 if the respondent lives in the corresponding number of the city district; 0 otherwise	AORI survey data
Albanians	Dichotomous var. that takes a value of 1 if the respondent declares that most of the foreigners living in their city district are from Albania; 0 otherwise.	AORI survey data
Pakistanis	Dichotomous var. that takes a value of 1 if the respondent declares that most of the foreigners living in their city district are from Pakistan; 0 otherwise.	AORI survey data

continued on the next page ...

Table A.1: Description of the variables – continued

Variable name	Variable description	Source
Africans	Dichotomous var. that takes a value of 1 if the respondent declares that most of the foreigners living in their city district are from Africa; 0 otherwise.	AORI survey data
Filipinos	Dichotomous var. that takes a value of 1 if the respondent declares that most of the foreigners living in their city district are from the Philippines; 0 otherwise.	AORI survey data
Syrians	Dichotomous var. that takes a value of 1 if the respondent declares that most of the foreigners living in their city district are from Syria; 0 otherwise.	AORI survey data

Table A.2: Correlation matrix of the variables of interest

	Perception of refugees' integration	Refugees' perceived presence	Refugees' actual presence	Migrants' perceived presence	Perception that foreigners cause problems
Perception of refugees' integration	1				
Refugees' perceived presence	-0.0640*	1			
Refugees' actual presence	0.0094	0.2058*	1		
Migrants' perceived presence	-0.1371*	0.3891*	0.2313*	1	
Perception that foreigners cause problems	-0.2467*	0.3994*	0.1470*	0.4149*	1

Note. $^*p < 0.05$.

Table A.3: Correlation matrix of the variables of interest transformed

	Perception of refugees' integration	Refugees' perceived presence	Refugees' actual presence	New migrants' perceived presence	Perception that foreigners cause problems 1	Perception that foreigners cause problems 2	Perception that foreigners cause problems 3	Perception that foreigners cause problems 4
Perception of refugees' integration	1							
Refugees' perceived presence	-0.0640*	1						
Refugees' actual presence	0.0094	0.2058*	1					
New migrants' perceived presence	-0.0768*	0.1969*	0.0767*	1				
Perception that foreigners cause problems 1	0.2145*	-0.3213*	-0.1218*	-0.1089*	1			
Perception that foreigners cause problems 2	-0.0463*	0.0418	0.0349	-0.0476*	-0.6452*	1		

continued on next page ...

Table A.3: Correlation matrix of the variables of interest transformed – continued

	Perception of refugees' integration	Refugees' perceived presence	Refugees' actual presence	New migrants' perceived presence	Perception that foreigners cause problems 1	Perception that foreigners cause problems 2	Perception that foreigners cause problems 3	Perception that foreigners cause problems 4
Perception that foreigners cause problems 3	-0.1553*	0.2294*	0.0454*	0.0841*	-0.4207*	-0.2418*	1	
Perception that foreigners cause problems 4	-0.1367*	0.2640*	0.1211*	0.1901*	-0.2779*	-0.1598*	-0.1042*	1

Note: $^*p < 0.05$.

Table A.4: Residents' perception of refugees' integration potential, Model 1 marginal effects

Dependent: Perception of refugees' integration	Cannot be integrated pr(y = 1)	Depends	Can be integrated pr(y = 1)
Refugees' perceived presence	-0.0223**	0.00313**	0.0192**
	(0.0111)	(0.00158)	(0.00959)
Perception that foreigners cause problems	0.122***	-0.0171***	-0.105***
	(0.0125)	(0.00205)	(0.0111)
Migrants' perceived presence	0.000724	-0.000102	-0.000623
	(0.000448)	(0.0000633)	(0.000385)
Unable to distinguish between migrants/refugees	0.0477	-0.00669	-0.0410
	(0.0358)	(0.00505)	(0.0308)
Gender	0.0106	-0.00149	-0.00912
	(0.0198)	(0.00278)	(0.0170)
Age	0.156***	-0.0219***	-0.135***
	(0.0445)	(0.00653)	(0.0383)
Age2	-0.0166***	0.00232***	0.0142***
	(0.00567)	(0.000820)	(0.00488)
Education	-0.0450***	0.00632***	0.0387***
	(0.0163)	(0.00233)	(0.0141)
Married with children over18	0.0151	-0.00212	-0.0130
	(0.0226)	(0.00318)	(0.0194)
Married with children under 18	0.0307	-0.00430	-0.0264
	(0.0316)	(0.00445)	(0.0272)
Unemployed	-0.00832	0.00117	0.00715
	(0.0330)	(0.00463)	(0.0284)
Inactive	0.0286	-0.00402	-0.0246
	(0.0298)	(0.00420)	(0.0257)
Income adequacy	-0.0203*	0.00285*	0.0175*
	(0.0116)	(0.00164)	(0.00995)
Born in Greece	0.103*	-0.0145*	-0.0886*
	(0.0598)	(0.00847)	(0.0515)
Political self-placement (left)	-0.224***	0.0314***	0.192***
	(0.0253)	(0.00452)	(0.0218)
Political self-placement (centre)	-0.0371*	0.00520*	0.0319*
	(0.022)	(0.003)	(0.019)
Perception that foreigners cause problems	0.122***	-0.0171***	-0.105***
	(0.0125)	(0.00205)	(0.0111)
N	2164	2164	2164
pseudo R^2			
AIC	.	.	.

continued on next page ...

Table A.4: Residents' perception of refugees' integration potential, Model 1 marginal effects – continued

Dependent: Perception of refugees' integration	Cannot be integrated pr(y = 1)	Depends	Can be integrated pr(y = 1)
BIC	.	.	.

Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.5: Residents' perception of refugees' integration potential, Model 2 marginal effects

Dependent: Perception of refugees' integration	Cannot be integrated pr(y = 1)	Depends	Can be integrated pr(y = 1)
Refugees' actual presence	-0.0000939**	0.0000136**	0.0000803**
	(0.0000457)	(0.00000668)	(0.0000391)
Perception that foreigners cause problems	0.117***	-0.0169***	-0.0996***
	(0.0113)	(0.00190)	(0.0100)
Migrants' perceived presence	0.000679*	-0.0000985*	-0.000581*
	(0.000409)	(0.0000597)	(0.000350)
Unable to distinguish between migrants/refugees	0.0338	-0.00490	-0.0289
	(0.0301)	(0.00438)	(0.0258)
Gender	0.0162	-0.00235	-0.0139
	(0.0186)	(0.00271)	(0.0159)
Age	0.145***	-0.0211***	-0.124***
	(0.0426)	(0.00641)	(0.0364)
Age^2	-0.0149***	0.00216***	0.0127***
	(0.00540)	(0.000804)	(0.00462)
Education	-0.0381**	0.00553**	0.0326**
	(0.0152)	(0.00224)	(0.0131)
Married with children over18	0.00458	-0.000665	-0.00392
	(0.0212)	(0.00307)	(0.0181)
Married with children under 18	0.0291	-0.00422	-0.0249
	(0.0303)	(0.00441)	(0.0259)
Unemployed	0.00599	-0.000868	-0.00512
	(0.0311)	(0.00451)	(0.0266)
Inactive	0.0441	-0.00639	-0.0377
	(0.0282)	(0.00412)	(0.0241)
Income adequacy	-0.0252**	0.00366**	0.0216**
	(0.0109)	(0.00160)	(0.00933)
Born in Greece	0.0879	-0.0127	-0.0751
	(0.0564)	(0.00824)	(0.0482)
Political self-placement (left)	-0.227***	0.0329***	0.194***
	(0.0239)	(0.00439)	(0.0205)
Political self-placement (centre)	-0.0416**	0.00604**	0.0356**
	(0.0205)	(0.00299)	(0.0176)
N	2433	2433	2433
pseudo R^2			
AIC	.	.	.
BIC	.	.	.

Standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: Residents' perception of refugees' integration potential with dummy variables

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
Refugees' perceived presence (ref. category: none)		
A few	0.272**	
	(0.108)	
Some	0.366***	
	(0.135)	
Many	0.0402	
	(0.195)	
Refugees' actual presence (ref. category: 50%)		
2%		-0.157
		(0.183)
8%		-0.225
		(0.144)
8%		-0.0551
		(0.151)
9%		-0.405***
		(0.138)
11%		-0.284*
		(0.149)
13%		-0.173
		(0.166)
Perception that foreigners cause problems (ref. category: none)		
A few	-0.571***	-0.536***
	(0.107)	(0.100)
Some	-1.273***	-1.206***
	(0.163)	(0.148)
Many	-1.555***	-1.559***
	(0.257)	(0.237)
Migrants' perceived presence	-0.00240	-0.00341*
	(0.00210)	(0.00193)
Unable to distinguish between migrants/refugees	-0.208	-0.158
	(0.167)	(0.140)
Gender	-0.0545	-0.0684
	(0.0921)	(0.0867)
Age (ref. category: 18-24)		
25-34	0.102	0.252
	(0.327)	(0.313)
35-44	-0.297	-0.226
	(0.278)	(0.265)
45-54	-0.555***	-0.530***
	(0.211)	(0.202)
55-64	-0.275*	-0.196
	(0.142)	(0.134)
Age^2	-0.0160*	-0.0160*
	(0.00854)	(0.00816)
Education		
Primary	-0.680	-0.565
	(0.948)	(0.908)
Secondary	-0.221	-0.208
	(0.934)	(0.896)
Tertiary	-0.0957	-0.0816
	(0.935)	(0.897)
Married with children over18	-0.0826	-0.0234
	(0.106)	(0.0991)
Married with children under 18	-0.135	-0.105
	(0.153)	(0.146)
Unemployed	0.0276	-0.0453
	(0.154)	(0.145)
Inactive	-0.138	-0.202
	(0.141)	(0.133)
Income adequacy (ref. category: making ends meet)		
Facing great difficulties	-0.211*	-0.281
	(0.121)	(0.251)
Facing difficulties	-0.0493	-0.134
	(0.119)	(0.250)

continued on next page ...

Table A.6: Residents' perception of refugees' integration potential with dummy variables – continued

Dependent: Perception of refugees' integration	Model 1 With perceived refugee presence	Model 2 With actual refugee presence
Living comfortably	-0.0561	
	(0.267)	
Born in Greece	-0.493*	-0.459*
	(0.280)	(0.263)
Political self-placement (left)	1.038***	1.042***
	(0.125)	(0.118)
Political self-placement (centre)	0.175*	0.205**
N	2164	2433
pseudo R^2	0.075	0.074
AIC	3870.3	4368.4
BIC	4035.0	4553.9

Standard errors in parentheses * p < 0.10, ** p < 0.05, *** p < 0.01

REGION
The Journal of ERSA
Powered by WU

Volume 12, Number 1, 2025, 73–101
DOI: 10.18335/region.v12i1.562

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

Spatial network analysis

Carmen Cabrera[1]

[1] University of Liverpool, Liverpool, United Kingdom

Received: 27 August 2024/Accepted: 14 February 2025

Abstract. The impact of current changes in technology, society, and the environment demand innovative research approaches that can accurately capture the increasing complexity of modern geographic data. To this end, this article provides an introduction to the use and implementation of spatial network analysis for the study of systems where geography matters. Presented as a computational notebook, the article offers practical, hands-on learning resources in R, assuming no prior knowledge from readers. Using a network of road links between African cities as case study, readers will be guided through key concepts and applications of spatial network analysis. The article highlights the relevance of network thinking in addressing contemporary geographic challenges.

1 Introduction

Sweeping transformations in technology, society and the environment are disrupting the status quo, driving innovation and challenging our understanding of the world. Among research fields, geography is significantly impacted by these changes due to its all-encompassing nature (Hartshorne 1939). The increased availability of data, the rise of Artificial Intelligence (AI), the improvement and deployment of urban sensors, the urbanisation and counter-urbanisation trends, the global spread of COVID-19, the evolving human mobility patterns, and the growing economic impact of natural disasters and climate change are just a few of the pressing issues that analysts interested in the study of geographic data are currently addressing.

Quantitative research concerning these new realities calls, more than ever, for approaches that embrace the interconnectedness and contextuality of the various elements involved in driving change. For example, a process like urbanisation is not just about the growth of urban populations but it also involves economic, social, and environmental dimensions and their interaction (Batty 2007, Bettencourt 2021). Similarly, climate change extends far beyond the realm of weather, with impacts on ecosystems, communities, financial markets and even politics (Lawrence et al. 2020). Responses to many of the emerging disruptions are additionally characterised by having ill-defined or malleable goals (e.g. achieving "sustainable urbanisation"), and by the presence of not only interdependent, but sometimes conflicting elements (e.g. designing a pandemic response while ensuring economic stability) (Miller et al. 2021). Therefore, the study of geographic data must adopt holistic frameworks and methodologies that regard the current real-world issues as far-reaching, wicked problems – as defined by (Rittel, Webber 1973) – embedded in complex systems (Miller 2017).

Complex systems are often organised as networked sets of elements which are embedded in space (Barthélemy 2011). These network structures, where space is relevant and where

topology alone does not contain all the information, are ubiquitous in the study of geographic data. Examples include transport networks (e.g. the subway network in a city, the network of flight connections between airports, the street network); mobility networks (e.g. the migration network between different countries, a network of commuting flows between the neighbourhoods within an urban area) or social and contact networks (e.g. the network of Facebook friends in different parts of the world). Therefore, we argue that network analysis remains one of the most significant and persistent research areas relating to the analysis of geographic data (Curtin 2007).

With roots in the natural sciences, the study of networks and complex systems explores how interactions among the individual give rise to larger, macro-level structures (Flake 1998). At the same time, network analysis provides powerful tools to investigate micro-level structures and properties, such as individual nodes and their local interactions, while embedding them within the broader context of the system. In this way, network analysis serves as a bridge between the "micro" and the "macro", offering insights into how local behaviours shape overarching patterns and vice versa.

Networks and complex systems thinking has traditionally focused on "extracting and abstracting", contrasting with the pre-disposition of geographers to "contextualise and specify" (O'Sullivan, Manson 2015). As a result, the use of complex systems approaches in geography and social sciences has faced criticism in the past for allegedly failing to capture human-driven qualitative change or power-dynamics, and neglecting substantive domain expertise (Uitermark, van Meeteren 2021, Franklin 2020). However, in recent times, with the rise of technological and socioeconomic transformations, alongside increasing data availability, there are numerous incentives to integrate complex systems approaches into geography (Nelson et al. 2025). These approaches should move away from sterile, often-critiqued conceptualisations of people and places, and instead, focus on data-driven analyses that consider the heterogeneity and spatial context as key drivers of social behaviours (Miller et al. 2021).

This article aims at providing an introduction of key concepts, definitions and applications of network analysis for the study of geographic data. Network analysis has a theoretical foundation in graph theory and topology (Curtin 2007). Rather than focusing on these foundational approaches to spatial networks, the focus here is on the practical implementation for data analysis and modelling. The article takes the form of a computational notebook, inviting the reader to follow a hands-on approach through a series of computational examples in R. These examples highlight the value of networks thinking in geography and spatial analysis by presenting applications based on a real-world network within a two-dimensional space.

The article is structured in the following way. The sections "Computational environment", "Data" and "Basic conceptual intuition" provide background and instructions to set up the necessary software to run the computational notebook. The body of the article can be found in the "Application" section, which includes learning resources for network construction and visualisation, network metrics, community detection and analysis of network robustness. Finally, there is a "Conclusion" section synthesising the learning outcomes.

2 Computational environment

This computational notebook is designed for reproducibility, so you can obtain consistent results by downloading and re-running without any modifications. To this end, it is necessary to ensure that the machine used to run the code has all the relevant software packages and versions installed.

The notebook is written using Quarto, an open-source, R Markdown-like publishing system which supports the integration of text, code and visualisations within a single document. Quarto documents are saved with the extension .qmd but can also be rendered in various formats, such as PDF, HTML, Microsoft Word or others. To use Quarto, you need to install the appropriate distribution for your operating system, which can be downloaded from the Quarto website. For the creation of this notebook, the version 1.5.55 (Mac OS) was used.

You will need to use RStudio, which requires R to be installed, to open and work on the Quarto notebook. R can be downloaded from the Comprehensive R Archive Network (CRAN), where installers for various operating systems, including Windows, macOS, and Linux, are provided. This notebook has been created using the R 4.4.1 version for MacOS, Apple Silicon (M1-3). Once R is installed, it is possible to proceed with the installation of RStudio, which can be downloaded from the Posit website. This notebook was created using the macOS 12+ version of RStudio.

You can open the notebook saved as a .qmd file on RStudio once you have installed Quarto, R, and RStudio. You will need to install a recent distribution of TeX such as TinyTeX to properly render the .qmd file into a .pdf. This can be done by running the following command on the terminal: quarto install tinytex. You will then need to add the Quarto template used to create this manuscript for REGION. Instructions on how to do this can be found in the README section of the following GitHub repository: https://github.com/region-ersa/REGION/. Furthermore, to run the code included in the notebook, you will need to install some R extensions, known as packages, that will be useful for the applications explored here. The packages, as well as the specific version that was used during the creation of the notebook, that you need to install are:

Package	Version	Description
`igraph`	2.0.3	for network manipulation and analysis
`sf`	1.0.16	to handle spatial data
`tidyverse`	2.0.0	for data manipulation and visualisation
`ggplot2`	3.5.1	for data visualisation
`ggraph-2`	.2.1	for graph visualisation
`patchwork`	1.2.0	to arrange plots
`tidygraph`	1.3.1	for tidy data handling with graphs
`RColorBrewer`	1.1.3	for color palettes
`rnaturalearth`	1.0.1	for natural earth map data
`ggspatial`	1.1.9	for geospatial visualisation

Open RStudio to install any package. Write the following command on the console window, normally situated at the bottom left: `install.packages("name of package")`. Make sure you replace "name of package" by the actual name of the package that you want to install e.g. `install.packages("tidyverse")`. Then, press enter and repeat this process until you have installed all the packages in the list.

Once the packages are installed, you will need to load them in order to be able to use them. This can be done by running the code below:

[1]:
```
# Load necessary libraries
library(igraph)          # Network analysis
library(sf)              # Handling spatial data (simple features)
library(tidyverse)       # Data manipulation and visualisation
library(ggplot2)         # Creating graphics and visualisations
library(ggraph)          # Visualising network data with ggplot2
library(patchwork)       # Combining multiple ggplot2 plots
library(tidygraph)       # Graph manipulation in a tidy data framework
library(RColorBrewer)    # Creating color palettes for visualisations
library(rnaturalearth)   # Accessing natural Earth geospatial data
library(ggspatial)       # Adding spatial context to ggplot2 maps
```

[2]:
```
# Disable scientific notation globally for figures
options(scipen = 999)
```

3 Data

Here, you get to work with a network which represents the main cities in the African continent and the road connections between them. The nodes of this network are the cities with a population greater than 100,000 people, obtained from (Moriconi-Ebrard et al.

2016). The edges represent the road infrastructure linking pairs of cities, sourced from OpenStreetMap (OpenStreetMap Contributors 2023, Maier 2014), and include primary roads, highways, and trunk roads. Each edge is associated with a distance measure based on the length of the roads that it represents. Additionally, to accurately represent the road infrastructure, the network also includes nodes representing road intersections, which are necessary for describing the connectivity between cities. These nodes are labelled as "transport nodes" and help define possible routes between cities. Some transport nodes correspond to towns with less than 100,000 inhabitants, so they are labelled as attached to nearby cities. The urban network enables us to consider the existing roads in the continent and measure the travelling distance rather than the physical distance between cities. The constructed network is formed by 7,361 nodes (2,162 cities and 5,199 transport nodes) and 9,159 edges. More details on how the network was built can be found in (Prieto-Curiel et al. 2022).

The network is connected, meaning that it is possible to find a sequence of nodes and existing roads linking any pair of cities, and therefore, it is also possible to find the shortest road distance between any two cities and define it as the network distance. The network consists of 361,000 km of road infrastructure and connects 461 million people living in African cities, representing roughly 39% of the continent's population (Prieto Curiel et al. 2022).

4 Basic conceptual intuition

Networks are used as a tool to conceptualise real-life systems where a group of items displays connections between themselves, such as the friendships among members of a school year group, hyperlinks between websites, or what-eats-what relationships in an ecological community. A network (or a graph) consists of nodes (or vertices) and edges (or links), where the latter represent the connections between the nodes (Aldous, Wilson 2000). Networks generally represent the arrangement of connections between pairs of nodes, also known as the topological information of the systems they represent. Networks are considered spatial when their topology alone does not provide all the necessary information. In such networks, the nodes are located in a space equipped with a metric (Barthélemy 2011), typically Euclidean distance in two-dimensional space. The probability of a link between two nodes generally decreases as the distance between them increases. An important type of spatial networks are planar networks, where edges can be drawn without crossing each other in a two-dimensional plane. However, they can also be non-planar, such as an airline passenger network that connects airports through direct flights where the flight connections often overlap geographically (Barthelemy 2018).

The geometry of the nodes is fundamental in spatial networks. While nodes may have explicit geographic coordinates, edges may not always directly reflect physical distances or real-world geometry. For example, the edges do not account for the actual routes or distances travelled in the case of straight-line connections between cities in a transportation network. However, the spatial arrangement of nodes still plays a significant role in the structure of the network.

The analysis of spatial networks is therefore an essential tool for studying patterns in geographic data. Spatial networks are found in a variety of contexts, such as transport links between subway stations (Cabrera-Arnau et al. 2023), flight connections between airports (Guimerà, Amaral 2004), neighbouring relationships between geographical locations (Anselin 1988), street networks in cities (Barthelemy, Boeing 2024) or road networks connecting cities (Prieto Curiel et al. 2022, Strano et al. 2017), as is the case of the application that we demonstrate in this article.

5 Application

You will now learn how to implement spatial network analysis for the study of geographic data. As described in the section Data, all the steps in the implementation will be exemplified with a network of African roads.

5.1 *Creating a network from a data frame*

The data that specifies the nodes and edges of the African road network is stored in two csv files, one for nodes and one for edges. This data can be loaded in two data frames:

[3]:
```
# Define URLs for nodes and edges CSV files
url_nodes <- paste0(
  "https://github.com/CrmnCA/spatial-networks-region-data/",
  "raw/main/data/AfricaNetworkNodes.csv"
)

url_edges <- paste0(
  "https://github.com/CrmnCA/spatial-networks-region-data/",
  "raw/main/data/AfricaNetworkEdges.csv"
)

# Read the CSV file containing network nodes data from the URL
df_nodes <- read.csv(url_nodes)

# Read the CSV file containing network edges data from the URL
df_edges <- read.csv(url_edges)
```

To provide a clearer understanding of the data structure, here are the first few rows of each data frame:

[4]:
```
# Display the first few rows of the nodes data frame
head(df_nodes)
```

[4]:
```
  Agglomeration_ID   agglosName      x       y  Pop2015 ISO3  Region
1             2320        Cairo 31.324  30.130 22995802  EGY   North
2             5199        Lagos  3.316   6.668 11847635  NGA    West
3             7098      Onitsha  6.928   5.815  8530514  NGA    West
4             4220 Johannesburg 28.016 -26.050  8314220  ZAF   South
5             4858     Kinshasa 15.293  -4.408  7270000  COD Central
6             5331       Luanda 13.385  -8.924  6979211  AGO Central
```

The `df_nodes` data frame specifies the properties of nodes in the network, including a unique ID for each node (`Agglomeration_ID`) and attributes such as geographic coordinates (`x` and `y`) or population size (`Pop2015`).

[5]:
```
# Display the first few rows of the edges data frame
head(df_edges)
```

[5]:
```
  from      to          l        h      time    timeU  timeUCB border
1 8211    2333  4.2943815 motorway  2.576629 80.44702 80.44702      0
2 8211 1000559  1.7716116 motorway  1.062967 15.14826 15.14826      0
3 8211 1000567  5.4142666 motorway  3.248560 17.33386 17.33386      0
4 8211    5425  0.7988003  primary  1.198201 21.55530 21.55530      0
5 8211 1054396 50.6469094  primary 75.970364 90.05566 90.05566      0
6 8211    8208  8.2435851  primary 12.365378 36.19194 36.19194      0
```

The `df_edges` data frame defines the connections between nodes. Each row contains a pair of nodes (i.e., "from" and "to") that are linked by an edge, along with attributes associated with the edge, such as `timeUCB` for travel time. Due to excessive computational costs to calculate these values within the computational notebook, it is important to note that the distances and travel times for each edge have been precomputed and stored in `df_edges`.

You can create a network as an `igraph` object using these data frames. Specifically, this network will be undirected (`directed = FALSE`), meaning the edges have no direction and the connection between two nodes is bidirectional. In other words, you can travel from A to B or from B to A without any directionality being implied if there is an edge between nodes A and B.

[6]:
```
# Create a network 'g' from data frames 'df_edges' and 'df_nodes'
g <- graph_from_data_frame(
  d = df_edges,
  vertices = df_nodes,
  directed = FALSE
)
```

When printed, the structure of an `igraph` object provides a summary of the network, including the number of nodes (vertices), edges, and key attributes. For example, printing the `g` object yields:

[7]:
```
g
```

[7]:
```
IGRAPH 354be48 UN-- 7361 9159 --
+ attr: name (v/c), agglosName (v/c), x (v/n), y (v/n), Pop2015 (v/n),
| ISO3 (v/c), Region (v/c), l (e/n), h (e/c), time (e/n), timeU (e/n),
| timeUCB (e/n), border (e/n)
+ edges from 354be48 (vertex names):
 [1] 2333--8211    8211--1000559 8211--1000567 8211--5425    8211--1054396
 [6] 8211--8208    8211--1055432 3467--1001315 3467--4936    3467--1296977
[11] 2333--1000607 2333--1001157 2333--1011439 2333--1027090 2333--1068481
[16] 2333--1116973 2333--3833    7356--6261    7356--1000307 7356--1000311
[21] 7356--1000863 7356--1041480 7356--1134459 6261--5163    7254--1000789
[26] 7254--1067779 2938--5889    2938--1032852 2938--1055190 2938--1055917
+ ... omitted several edges
```

UN– indicates that the graph is undirected, with 7,361 nodes and 9,159 edges. The graph also includes vertex attributes (`name`, `agglosName`, `x`, `y`, etc.) and edge attributes (`l`, `h`, `time`, `timeU`, etc.).

You can examine the attributes of the nodes in the network, which are automatically derived from the column names in the `df_nodes` data frame. These attributes provide additional information about the nodes and are an essential part of working with `igraph` objects:

[8]:
```
# Retrieve the attribute names associated with nodes in the 'g' network
vertex_attr_names(g)
```

[8]:
```
[1] "name"       "agglosName" "x"          "y"          "Pop2015"
[6] "ISO3"       "Region"
```

Specifically, `name` is the ID of each node in the network, `agglosName` is the name of the city represented by the node and it is set to `road` if the node is a transport node. `x` and `y` represent the coordinates of each node, `Pop2015` is the population of the city nodes, `ISO3` is the code for the country that each node is situated in, `Region` represents the region within the African continent that each node is situated in, and `Between` and `degree` represent the betweenness centrality and the degree of each node in the network, which you will also compute below.

You can look particularly at the first few values of any node attribute, for example `Pop2015`:

[9]:
```
# Retrieve the first few node names from the 'g' network
head(V(g)$Pop2015)
```

[9]:
```
[1] 22995802 11847635  8530514  8314220  7270000  6979211
```

You can also obtain the names of the edge attributes, which are taken from the columns in the `df_edges` data frame:

[10]:
```
# Retrieve the attribute names associated with edges in the 'g' network
edge_attr_names(g)
```

[10]:
```
[1] "l"       "h"       "time"    "timeU"   "timeUCB" "border"
```

where `l` represents the length in kilometres by road segment and it considers curves, `h` is the type of edge (primary, highway, etc.), `time` is the estimated minutes to travel through the edge, considering different speeds for distinct types of road, `timeU` is also the estimated minutes to travel through the edge, but allowing extra time if the ends of the edge are urban nodes, `timeUCB` allows further extra time for edges that cross a border and `border` is a binary variable taking value 1 if an edge crosses a border and 0 otherwise.

For further reading on `igraph` and related network analysis tools in R, Kolaczyk, Csárdi (2014) offers a detailed introduction to `igraph` functionalities.

5.2 Visualising the African road network as a spatial network

The most basic network visualisation can be achieved with the `plot` function from `igraph`. While the arguments of this function allow for some degree of customisation, the `ggraph` library is integrated with `ggplot2`, a powerful and flexible tool for the creation of graphics which provides advanced control over aesthetics, layering, and themes.

We will load the shapes of the African countries as a spatial feature object, with the `ne_download` function before creating the visualization. This will be used as a basemap.

[11]:
```
# Download world map data with specified parameters
world <- ne_download(
  scale = "small",
  category = "cultural",
  type = "admin_0_countries",
  returnclass = "sf"
)

# Extract unique country ISO3 codes from `df_nodes` to match the network
target_countries <- unique(df_nodes$ISO3)

# Subset the world map data to include only the target countries
world_subset <- world[world$SOV_A3 %in% target_countries, ]
```

We will also need to set the position of the nodes by specifying their layout:

[12]:
```
custom_layout <- data.frame(
  name = df_nodes$agglosName,  # Node names from the graph
  x = df_nodes$x,              # Custom x-coordinates
  y = df_nodes$y               # Custom y-coordinates
)
```

Furthermore, the node size can be set to vary according to the population of the cities that they represent. City population sizes vary over several orders of magnitude, so rather than making the size of the nodes proportional to the population size, it is better to apply a scaling function to reduce the disparity in sizes:

[13]:
```
# Calculate and assign a 'size' attribute to nodes in the 'g' network
# Size is determined based on the population data of each node
V(g)$size <- 0.1 * (V(g)$Pop2015 / 80000)^3
```

Now, with a few modifications to the default plot in order to improve the appearance, including setting the size of nodes as a function of the population, the network is ready to be plotted.

[14]:
```
ggraph(as_tbl_graph(g), custom_layout) + # Basic graph plot
  geom_edge_link(
    color = "gray20",
    alpha = 0.9,
    aes(width = E(g)$l * 0.1)  # Custom edges
  ) +
```

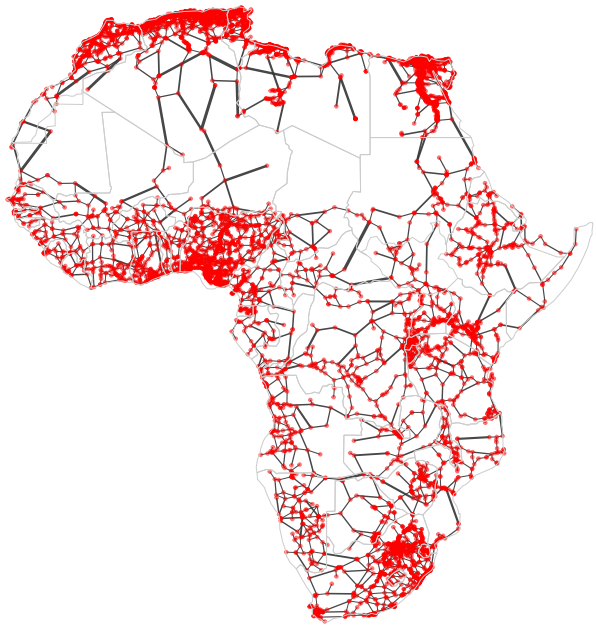


Figure 1: Visualisation of the African road network.

```
  scale_edge_width(range = c(.1, 0.7)) +  # Scale edge size
  geom_node_point(
    aes(color = "red", alpha = 0.8, size = V(g)$size)  # Custom nodes
  ) +
  scale_size_continuous(range = c(.3, 4)) +  # Scale node size
  scale_color_identity() +  # Scale node color
  theme(
    legend.position = "none",
    panel.background = element_rect(fill = NA, colour = NA)
  ) +
  geom_sf(
    data = world_subset,
    fill = NA,
    color = "gray80"  # Basic map plot
  )
```

[14]:
```
Output in Figure 1.
```

As an exercise, you may want to try to plot the default visualisation by simply running `plot(g)`. This demonstrates how simple modifications to the default plot can make a significant difference in the appearance of the outcome.

5.3 Network metrics

Network metrics are useful to obtain measurable insights into the network structure. They are also valuable as a way to characterise the network so that it can be compared to other networks.

5.3.1 Density

The density of a network refers to the proportion of existing edges over all possible edges. In a network with n nodes, the total number of possible edges is $n \times (n-1)/2$. A density equal to 1 corresponds to a situation where $n \times (n-1)/2$ edges are present. A network

with no edges at all would have density equal to 0. We can obtain the density of the African road network by running the following code:

[15]:
```
# Calculate edge density (excluding loops)
dens <- edge_density(g, loops = FALSE)
dens
```

[15]:
```
[1] 0.0003381142
```

The edge density for the African road network is approximately 0.00034, giving an indication that the network is quite sparse, since out of all possible edges, only 0.034% are present.

Beyond suggesting limited connectivity, low edge density also translates into challenges in transportation infrastructure, particularly in regions where road networks are sparse and fragmented due to geographical, economic, and historical factors. From an economic standpoint, low network density can hinder economic integration and regional trade by limiting connections between key locations (Linard et al. 2012). Improving connectivity in such sparse networks can reduce transportation costs, improve access to markets, and create greater economic opportunities, highlighting the importance of road network density for fostering economic development (Prieto Curiel et al. 2022).

5.3.2 Reciprocity

The reciprocity in a directed network is the proportion of reciprocated connections between nodes (i.e. number of pairs of nodes with edges in both directions) from all the existing edges.

[16]:
```
# Calculate the reciprocity of the edges in the 'g' network
reciprocity(g)
```

[16]:
```
[1] 1
```

Every edge is inherently bidirectional in an undirected graph, meaning that there is always a corresponding edge from B to A if there is an edge between node A and node B. Therefore, the reciprocity of an undirected graph is naturally 1, as all connections are reciprocated by definition.

5.3.3 Distances

A path in a network between node A and node B is a sequence of edges joining distinct nodes, starting at node A and ending at node B. Each node in the path is visited only once. In the case of a directed path, all edges must align with the specified direction, ensuring the path follows the network's directional flow.

The length of a path between nodes A and B is generally defined as the number of edges forming this path. The shortest path is the minimum count of edges present to travel from A to B. The path length can be also be defined in alternative ways. For example, the path length can be defined as the sum of the weights of the edges forming a path if the edges are weighted.

We can use the function `shortest_paths()` to find the shortest path between a given pair of nodes, taking into account the geographic road length associated with the edges. For example, between Cairo and Lagos, we can apply `shortest_paths()`, setting `weights` to `df_edges$l`, and store the output in a dataframe called `df_shortest_path`.

[17]:
```
# Calculate the shortest path between "Cairo" and "Lagos"
# Edge length is used as weight
df_shortest_path <- shortest_paths(
  g,
  from = V(g)$agglosName == "Cairo",
  to = V(g)$agglosName == "Lagos",
  predecessors = FALSE,
```

```
  weights = df_edges$l,
  output = "both"
)
```

In this dataframe, the field `epath` stores the edges of the shortest path as a one-element list. We can extract the values of this list as the edge IDs, which we then use to compute the geographic road length associated with the shortest path between the two network nodes.

[18]:
```
# Get the edge path indices from 'df_shortest_path'
idx <- df_shortest_path$epath[[1]]

length(idx)
```

[18]:
```
[1] 143
```

[19]:
```
# Get the lengths of edges along the path
lengths_epath <- edge_attr(g, "l", idx)

# Calculate the total length of the path
sum(lengths_epath)
```

[19]:
```
[1] 6084.359
```

We find that the network length associated with the shortest path is 143, while the geographic road length associated with this path is 6,084.359 km. This road distance, derived from the network representation of the African road network, can be compared with distances provided by routing services. For instance, Google Maps estimates the road transport distance between these two cities to be 6,431 km, which represents only a 5.7% difference from the value obtained here."

The diameter of a network is the longest shortest path between any pair of nodes, measured in terms of network distance. In the case of the African road network, we can incorporate the geographic road length as edge weights. Using these weights, the diameter reflects the greatest road length between any two nodes in the network.

[20]:
```
# Calculate the diameter of the network
diameter(g, directed = FALSE, weights = df_edges$l)
```

[20]:
```
[1] 11987.06
```

We obtain a diameter of 11,987.06 km, which roughly corresponds to the distance between the northernmost and southernmost points of the African continent.

The mean distance is the average length of all shortest paths in the network. The mean distance will always be smaller or equal than the diameter.

[21]:
```
# Calculate the mean distance of the network
mean_distance(g, directed = FALSE, weights = NULL)
```

[21]:
```
[1] 55.8602
```

[22]:
```
# Calculate the mean distance of the network
mean_distance(g, directed = FALSE, weights = df_edges$l)
```

[22]:
```
[1] 4878.73
```

Here we see that the average network distance between any given pair of nodes is 55.86 edges, while the geographic road distance between any given pair of nodes is 4,878.73 km.

5.3.4 Centrality

Centrality metrics assign scores to nodes, and sometimes edges, according to their position within a network. These metrics can be used to identify the most influential or important

nodes. In addition to measuring the prominence of individual nodes, centrality metrics also provide insight into the overall structure of the network. We can understand how local interactions at the node level contribute to global patterns and behaviours within the network by examining the distribution of centrality scores. Additionally, the degree distribution can reveal whether a network is highly centralised, with a few nodes having disproportionately high centrality scores, or whether it is more balanced, with nodes having similar centrality scores. This link between micro-level node properties and the broader macro-level structure highlights the role of network analysis in connecting individual node dynamics with the overall system.

5.3.4.1 Degree The degree of a node is one of the simplest and most fundamental measures of centrality in a network. It is defined as the number of edges connected to the node. In directed networks, the degree can be further divided into the in-degree, which counts the number of edges directed towards a node, and the out-degree, which counts the number of edges originating from it. The `degree()` function enables the calculation of these measures for one or more nodes in a network. Users can specify whether they are interested in the total degree (combining in- and out-degrees), the in-degree, or the out-degree, depending on the focus of their analysis.

We can compute the degree of each node with the `degree` function since the African road network is undirected.

[23]:
```
# Compute degree of the nodes given by v belonging to network g
deg <- degree(g, v = V(g))
```

We produce a histogram to visualise the results.

[24]:
```
# Produce a histogram showing frequency of nodes with certain degree
hist(deg,
     breaks = 50,
     main = "Distribution of degree",
     xlab = "Degree",
     ylab = "Frequency")
```

[24]:
```
Output in Figure 2.
```

We observe that most nodes have degree 3. Nodes of degree 1 are terminal nodes. Nodes of degree 2 are relatively less common than those of degree 1 and 3. This is likely due to the method used to build the network, where all the transport nodes of degree 2 are eliminated in order to simplify the network. It is relatively rare to find any nodes beyond degree 4. From the histogram, we see the maximum degree observed in the network is 13. Below, we obtain the name of the node with the maximum degree as well as the value of the degree (13).

[25]:
```
# Get names of nodes with the highest degree in the 'g' network
V(g)$agglosName[
  degree(g) == max(degree(g))
]
```

[25]:
```
[1] "Duduza Central"
```

[26]:
```
# Get IDs of nodes with the highest degree in the 'g' network
highest_degree_node_names <- V(g)$name[
  degree(g) == max(degree(g))
]
# Calculate the degree of nodes with the highest degree
degree(
  g,
  v = highest_degree_node_names
)
```

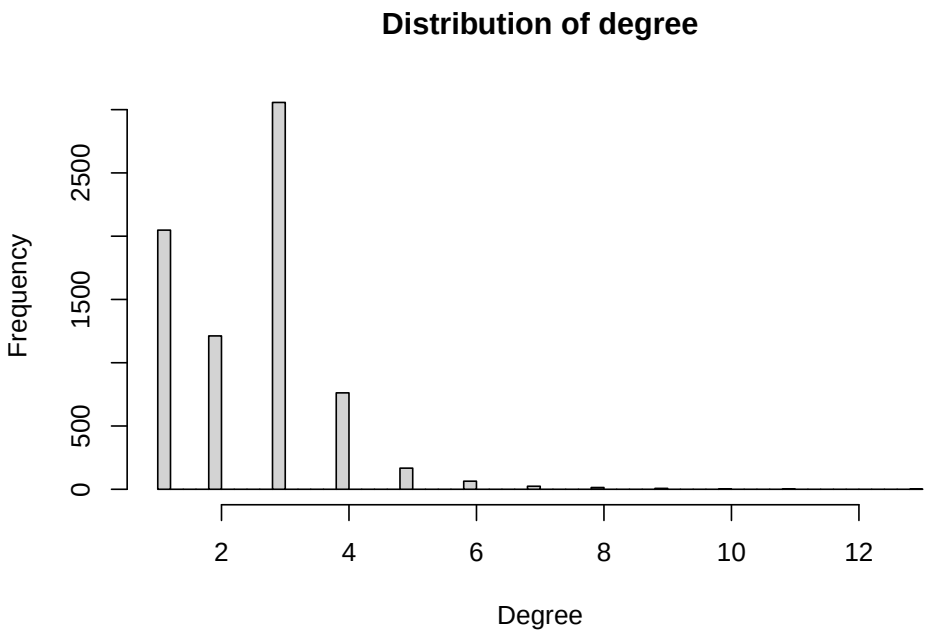


Figure 2: Histogram of node degree.

[26]:
```
2896
  13
```

The degree of 13 for Duduza means that it is directly connected to 13 neighbouring nodes (cities or intersections) in the road network. This degree indicates that Duduza serves as a relatively well-connected hub in the network.

We can also measure the weighted degree of a node. This is known as the strength of a node and it is computed as the sum of edge weights linked to adjacent nodes. Both the degree and strength are considered to be centrality metrics.

5.3.4.2 Closeness centrality Closeness centrality is a measure of the shortest path between a node and all the other nodes, measured in terms of network distance. It is calculated as the inverse of the average shortest path between a node and all other nodes. The closer this value is to 1, the more central the node is in the network. A value of 0 indicates an isolated node. You can use the `closeness()` function to compute closeness centrality for a network. The calculation assumes unweighted edges when setting `weights = NULL`, meaning the shortest paths are computed without accounting for edge weights such as geographic road lengths. Below, we compute the closeness centrality for a network using unweighted edges and visualise the results in a histogram to examine the distribution.

[27]:
```
# Calculate the closeness centrality for each node (unweighted edges)
close_centr <- closeness(g, weights = NULL)

# Create a histogram of closeness centrality
hist(close_centr,
     breaks = 50,
     main = "Distribution of closeness centrality",
     xlab = "Closeness centrality",
     ylab = "Frequency")
```

[27]:
```
Output in Figure 3.
```

The resulting distribution is bimodal. This distribution reflects the effect of geography in determining the overall connectivity structure of the African road network, whereby

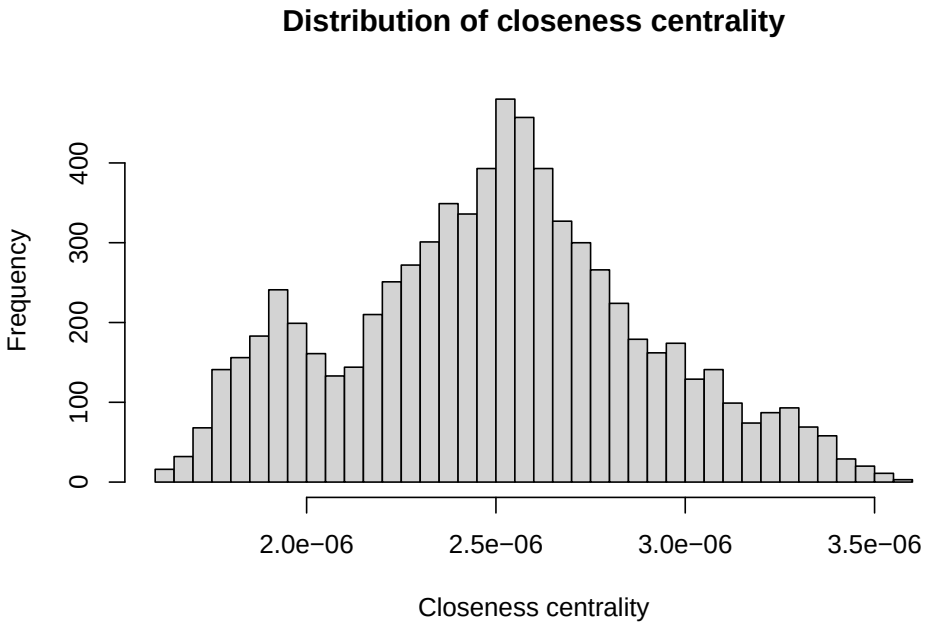


Figure 3: Histogram of unweighted closeness centrality.

the Sahara desert practically separates the network into two groups of nodes. Nodes within each of the groups, corresponding to cities either North or South of the desert, are much closer to each other than to the nodes belonging to the other group. Notably, the closeness centrality values are very low. This is due to the sparsity of the network, as closeness centrality is inversely related to the shortest path distances between nodes. Many nodes are far apart in a sparse network, resulting in very small closeness centrality values.

Closeness centrality can also be computed using weighted edges. In this case, geographic road length can be used as a weight to measure the physical proximity of each node to all other nodes in the network. The resulting distribution can be visualised using a histogram.

[28]:
```
# Calculate the closeness centrality for each node (weighted edges)
close_centr_weight <- closeness(g, weights = df_edges$l)

# Create a histogram of closeness centrality
hist(close_centr_weight,
     breaks = 50,
     main = "Distribution of weighted closeness centrality",
     xlab = "Closeness centrality (weighted)",
     ylab = "Frequency")
```

[28]:
```
Output in Figure 4.
```

Notably, the shape of the distribution changes when considering weighted or unweighted edges, demonstrating the importance of this choice in the outcome.

5.3.4.3 Betwenness centrality Similarly, betweenness centrality is a measure of the number of shortest paths going through a node. High values of betweenness centrality indicate that the corresponding nodes play an important role in the overall connectivity of the network. Betweenness can also be computed for edges. We compute the betweenness centrality for all nodes using the function `betweeness()` and represent it as a histogram. We do this using unweighted edges, so the computation of betweenness considers only network distances but not geogrpahic distances.

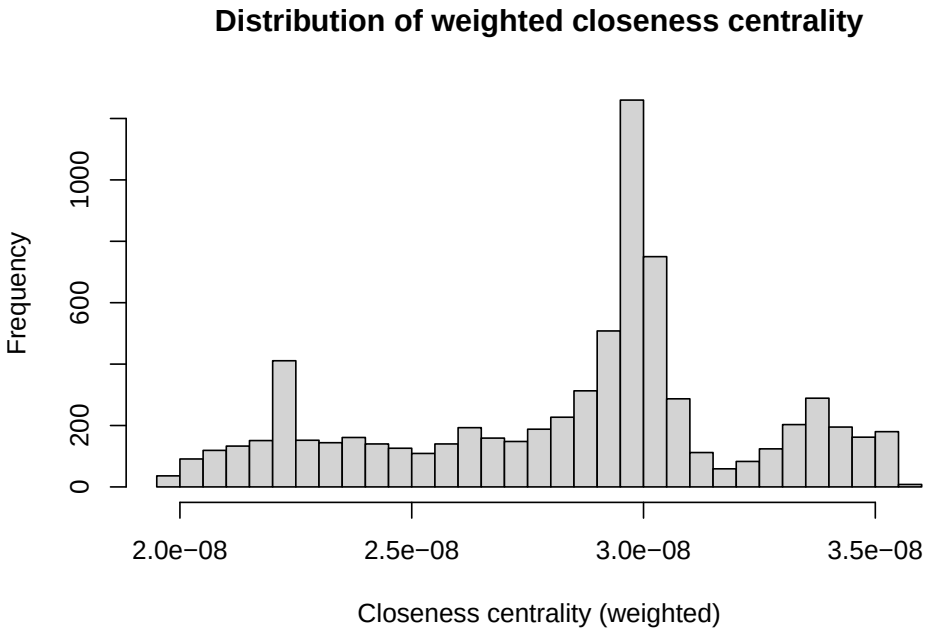


Figure 4: Histogram of weighted closeness centrality.

[29]:
```
# Calculate the betweenness centrality for each node in the g network
between_centr <- betweenness(g, v = V(g), directed = FALSE,
     weights = NULL)

# Create a histogram of betweenness centrality
hist(between_centr,
     breaks = 50,
     main = "Distribution of unweighted betweenness centrality",
     xlab = "Betweenness centrality",
     ylab = "Frequency")
```

[29]:
```
Output in Figure 5.
```

The resulting distribution shows that most nodes have low values of betweenness centrality and very few of them have large values. This means that there is a small number of nodes that are crucial to ensure the existence of a shortest path between any given pair of nodes. Without these key nodes, distances within the network could become larger or even infinite, since the network could be broken into isolated components. The emergence of such skewed distribution of the nodes betweenness centrality is due to the specific geographic features of the African road network, where once again, connections between cities North and South of the Sahara are facilitated by a small number of nodes with high values of betweenness centrality.

5.3.5 Assortativity

The assortativity coefficient quantifies the tendency of nodes in a network to connect to others with similar or dissimilar attributes. This metric can be calculated based on any standard node property, such as degree, or a custom attribute (e.g., population size). Mathematically, it is defined as the Pearson correlation coefficient of the specified attribute between pairs of connected nodes, with values ranging from -1 to 1. Positive assortativity indicates that nodes are more likely to connect with others having similar attributes. Negative assortativity suggests that nodes tend to connect with others having dissimilar attributes. In `igraph`, the `assortativity()` function computes the assortativity coefficient for a specified attribute, while the `assortativity_degree()`

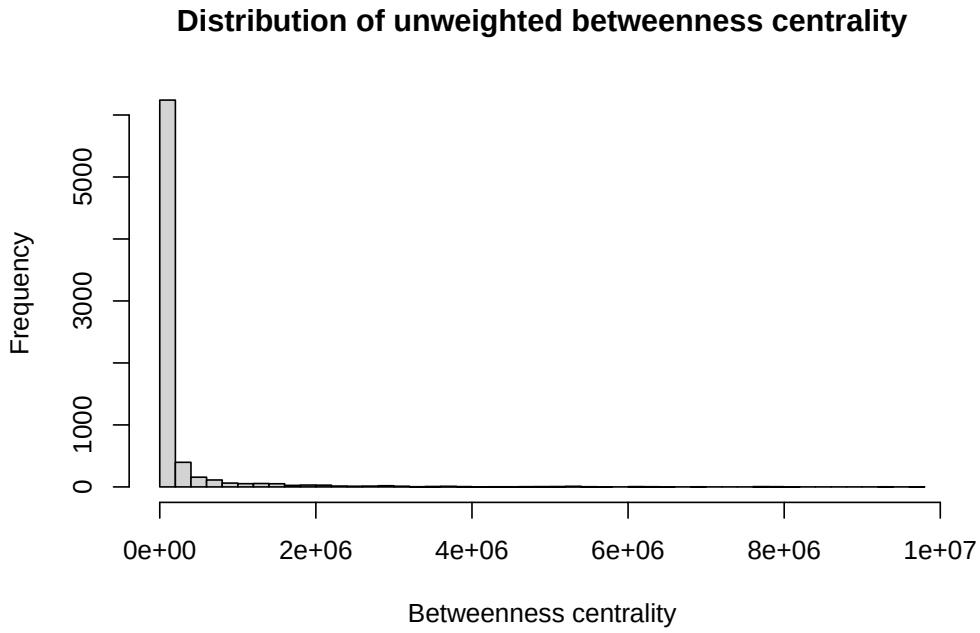


Figure 5: Histogram of betweenness centrality.

function specifically evaluates the assortativity based on node degrees.

Below is the code for computing assortativity of a custom network attribute such as the population size.

[30]:
```
assortativity(
  g,
  V(g)$Pop2015,
  directed = FALSE
)
```

[30]:
```
[1] 0.0006432013
```

We use the `assortativity_degree()` function to compute the degree assortativity:

[31]:
```
# Compute assortativity for degree
assortativity_degree(g, directed = FALSE)
```

[31]:
```
[1] -9.080965e-05
```

Both `assortativity()` and `assortativity_degree()` return values close to zero, indicating the absence of a strong tendency for nodes to connect either to similar or dissimilar nodes.

To better understand how degree assortativity manifests in the network, we can visualise the joint degree distribution, which shows the frequency of edges connecting nodes with specific degree pairs:

[32]:
```
# Extract node degrees
deg <- degree(g)

# Compute degree pairs for all edges
edge_degree_pairs <- t(sapply(E(g), function(e) {
  ends(g, e, names = FALSE) %>%
    sapply(function(v) deg[v])
}))
```

```
# Convert to data frame for visualisation
joint_degree_df <- as.data.frame(edge_degree_pairs)
colnames(joint_degree_df) <- c("Degree_1", "Degree_2")

# Aggregate counts for each degree pair
joint_degree_counts <- joint_degree_df %>%
  group_by(Degree_1, Degree_2) %>%
  summarise(Frequency = n(), .groups = "drop")

# Heatmap visualisation
ggplot(joint_degree_counts, aes(x = Degree_1,
                                y = Degree_2,
                                fill = Frequency)) +
  geom_tile(color = "white") +
  scale_fill_gradientn(
        colors = c("white", "lightblue", "blue", "darkblue"),
        trans = "log10",
        name = "Frequency") +
  labs(title = "Joint Degree Distribution",
        x = "Degree of Node 1",
        y = "Degree of Node 2",
        fill = "Log-scaled Frequency") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12),
        legend.position = "right") +
  scale_x_continuous(
    breaks = seq(1, max(joint_degree_counts$Degree_1), by = 1)
  ) +
  scale_y_continuous(
    breaks = seq(1, max(joint_degree_counts$Degree_2), by = 1)
  ) +
  coord_cartesian(expand = FALSE) +
  geom_hline(yintercept = seq(0.5, max(joint_degree_counts$Degree_2) +
      0.5,by = 1),
    color = "grey90", linetype = "solid") +
  geom_vline(xintercept = seq(0.5, max(joint_degree_counts$Degree_1) +
      0.5,by = 1),
    color = "grey90", linetype = "solid")
```

[32]:
```
Output in Figure 6.
```

Since the `assortativity()` and `assortativity_degree()` functions in `igraph` provide only the assortativity coefficient, a permutation test can be implemented to assess the statistical significance of the assortativity coefficient. A permutation test compares the observed assortativity coefficient to a distribution of coefficients from randomly shuffled networks. This helps determine whether the observed value is significantly different from what could arise by chance.

Here is how to compute a p-value using a permutation test:

[33]:
```
# Observed degree assortativity coefficient
observed_assortativity <- assortativity_degree(g)

# Set seed for reproducibility
set.seed(123)
```

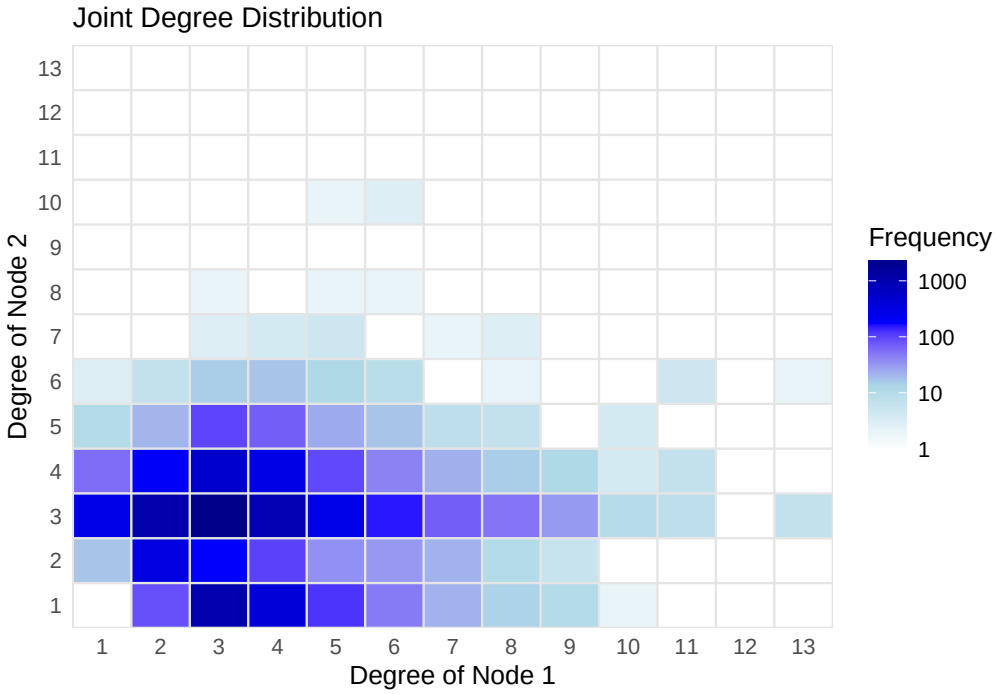


Figure 6: Joint degree distribution

```
n_permutations <- 1000

# Rewire edges and compute assortativity for random networks
random_assortativities <- replicate(n_permutations, {
  g_random <- rewire(g, keeping_degseq(niter = ecount(g) * 10))
  assortativity_degree(g_random)
})

# Compute p-value by comparing observed and random assortativities
p_value <- mean(
  abs(random_assortativities) >= abs(observed_assortativity)
)

p_value
```

```
[33]: [1] 0.997
```

We cannot reject, at the 95% confidence level, the hypothesis that the observed assortativity could arise from a random configuration, since p-value ≥ 0.05. If p-value was < 0.05, the observed assortativity would be unlikely to arise by chance, indicating significant assortativity.

5.4 Community detection

A network displays community structure if the nodes can be grouped into sets such that the nodes within each set are densely connected. For example, in the case of a social network formed by the students in a classroom, we would expect that small groups of friends form within the overall network, where relationships among members of a group are stronger than to everyone else in the classroom. Detecting or searching communities within in a network is a fundamental problem in network analysis, which has attracted much attention in the past decades (Fortunato, Newman 2022). While there are different methods to detect communities, below we review three of them which have been widely

used and studied. These are walktrap (Pons, Latapy 2005), edge-betweenness (Girvan, Newman 2002) and the Louvain method (Blondel et al. 2008).

Each of these methods emphasises different aspects of community structure, leading to varying outcomes. The walktrap method is particularly suited to identifying small, localised communities, making it useful in applications where fine-grained clusters are of interest, such as detecting close-knit friend groups or tightly interconnected modules in biological networks. The edge-betweenness method excels at uncovering well-separated communities, especially in networks where connections between groups are sparse or function as bottlenecks, such as departments in an organisation or transportation networks. The Louvain methdod is best for capturing broader, hierarchical structures and is especially effective for large-scale networks, where computational efficiency and an overview of the global structure are key.

The community structures they produce can differ significantly because these methods focus on different aspects of the network. For instance, walktrap or edge-betweenness may reveal smaller, more detailed clusters, whereas Louvain often detects larger, coarser groupings. It is important to choose a method thoughtfully based on the specific application and research question because the detected community structure can directly shape the interpretation of the network.

It is also worth noting that both walktrap and the Louvain method incorporate random elements in their algorithms. As a result, running these methods on the same network can yield slightly different community structures across iterations. To address this variability, it is recommended to run the algorithms multiple times and assess the stability of the detected communities using similarity metrics, such as the RAND index (Rand 1971), which quantify the agreement between different community partitions. Including such assessments can improve confidence in the robustness of the results and ensure reproducibility, especially for applications where stable community detection is relevant.

5.4.1 Walktrap

This algorithm relies on the concept of random walks on networks. Random walks are sequences of nodes, chosen by following a randomly chosen path. The underlying assumption of the walktrap method is that nodes encountered in a given random walk are more likely to be part of the same community.

The algorithm starts by treating each node as its own community. Then, it performs a series of short random walks on the network, where the length of these walks has to be specified by the user. After performing the random walks, the algorithm calculates a similarity measure between each pair of nodes. This measure is based on the idea that if two nodes are often encountered together during random walks, they are likely part of the same community. Nodes that have high similarity are merged into larger communities. This merging process is hierarchical and agglomerative, starting with individual nodes and progressively combining them. As communities are merged, the algorithm often aims to maximise a measure called modularity, which quantifies the strength of the division of the network into communities. High modularity indicates a good community structure, where more edges fall within communities than between communities. The process continues until the entire network is merged into a single community or until a stopping criterion, like a modularity threshold is met. The algorithm may return a hierarchical structure of communities depending on the implementation, allowing the user to explore different levels of granularity in the community structure.

The walktrap method is implemented in R via the `igraph` function `cluster_walktrap()`, with key parameters including the network of interest as an igraph object, the length of the random walks, and a membership parameter, which is a boolean variable indicating whether to calculate membership based on the highest modularity score (with True as the default). Below, we apply the walktrap method to the African road network and save the result in the variable `g_wt`:

[34]:
```
# Perform walktrap clustering on the graph
g_wt <- cluster_walktrap(graph = g, steps = 3, membership = TRUE)
```

We can get the membership of each node as well as the modularity score according to the solution based on random walks of length 3. We save the results with the name `member_g_wt` and `modularity_g_wt`:

[35]:
```
# Get the membership of clusters in the walktrap clustering
member_g_wt <- membership(g_wt)

# Calculate the modularity of the walktrap clustering
modularity_g_wt <- modularity(g_wt)
```

We can plot the network based on the found communities. We will plot the network as before, but color the nodes based on the communities found using the walktrap algorithm. The plot is included in Figure 7:

[36]:
```
plot_wt <- ggraph(as_tbl_graph(g), custom_layout) +  # Basic graph plot
  geom_edge_link(color = "gray20",
                 alpha = 0.9,
                 aes(width = E(g)$l * 0.1)) +
  scale_edge_width(range = c(.1, 0.7)) +  # Scale edge size
  geom_node_point(aes(color = member_g_wt,
                      size = V(g)$size,
                      alpha = 0.8)) +
  scale_size_continuous(range = c(.3, 4)) +  # Scale node size
  scale_color_identity() +  # Scale node color
  theme(legend.position = "none",
    panel.background = element_rect(fill = NA, colour = NA)) +
  geom_sf(data = world_subset,
          fill = NA,
          color = "gray80") +
  ggtitle("Walktrap method")
```

5.4.2 Edge betweenness community detection

The edge-betweenness community detection method identifies communities within a network by focusing on the edges that connect different communities. It works by progressively removing edges that act as bridges between groups of nodes, from higher to lower betweenness centrality. As high-betweenness edges are removed, the network breaks down into smaller, more cohesive subgroups or communities. Though it can be computationally intensive for large networks, this method is particularly useful for finding natural divisions within a network.

The algorithm starts by computing the betweenness centrality for all edges in the network. Edges with high betweenness are likely to be those that connect different communities. Then, the edge with the highest betweenness centrality is removed from the network. This step effectively "cuts" the bridge between communities. After removing the edge, the betweenness centrality for the remaining edges is recomputed since the removal of one edge may change the shortest paths in the network, affecting the betweenness centrality of other edges. Edges with the highest betweenness centrality keep being removed until all edges have been removed or until the network breaks down into the desired number of communities.

The edge betweenness community detection method is implemented in R via the `igraph` function `cluster_edge_betweenness()`. The key parameters are the network of interest as an igraph object and a membership parameter, a boolean variable indicating whether to calculate membership based on the highest modularity score (with True as the default). Below, we apply the edge betweenness method to the African road network and save the result in the variable `g_eb`:

[37]:
```
# Perform edge betweenness clustering on the graph
g_eb <- cluster_edge_betweenness(graph = g, membership = TRUE)
```

Once again, we can get the membership of each node according to the solution based on edge betweenness. We save the results with the name `member_g_eb`:

[38]:
```
# Get the membership of clusters in the edge betweenness clustering
member_g_eb <- membership(g_eb)
```

Then, we generate a plot of the results. We include the plot in Figure 7.

[39]:
```
plot_eb <- ggraph(as_tbl_graph(g), custom_layout) +  # Basic graph plot
  geom_edge_link(color = "gray20",
                 alpha = 0.9,aes(width = E(g)$l * 0.1)) +
  scale_edge_width(range = c(.1, 0.7)) +  # Scale edge size
  geom_node_point(aes(color = member_g_eb,
                      size = V(g)$size,
                      alpha = 0.8)) +
  scale_size_continuous(range = c(.3, 4)) +  # Scale node size
  scale_color_identity() +  # Scale node color
  theme(legend.position = "none",
        panel.background = element_rect(fill = NA, colour = NA)) +
  geom_sf(data = world_subset,
          fill = NA,color = "gray80") +
  ggtitle("Edge betweenness\n clustering")
```

5.4.3 Louvain method

The Louvain method of multi-level clustering works by finding communities in such a way that the modularity of the network is maximised. The algorithm works in two phases: first, each node starts in its own community, and nodes are iteratively moved to neighboring communities if the move increases modularity. This phase continues until no further improvement is possible. In the second phase, the network is compressed by treating each community found in the first phase as a single node, creating a new, smaller network. The two phases are then repeated on this simplified network, thus refining the community structure at each level. The process continues until modularity no longer increases, resulting in a hierarchical clustering that reflects the community structure within the network.

In R, the Louvain method is implemented via the `cluster_louvain()` function, where the arguments are the graph and the resolution. Higher resolution values will yield a larger number of smaller communities, while lower values will yield a smaller number of larger communities.

[40]:
```
# Perform Louvain clustering on the graph with a resolution of 1
g_mlc <- cluster_louvain(graph = g, resolution = 1)
```

We get the membership of each node according to the communities found by the Louvain's multi-level clustering method. The results are saved with the name `member_g_mlc`:

[41]:
```
# Get the membership of clusters in the Louvain clustering
member_g_mlc <- membership(g_mlc)
```

We can then generate a plot of the results. The plot is included in Figure 7:

[42]:
```
plot_mlc <- ggraph(as_tbl_graph(g), custom_layout) +  # Basic graph plot
  geom_edge_link(color = "gray20",
                 alpha = 0.9,
                 aes(width = E(g)$l * 0.1)) +
  scale_edge_width(range = c(.1, 0.7)) +  # Scale edge size
  geom_node_point(aes(color = member_g_mlc,
                      size = V(g)$size,
                      alpha = 0.8)) +
```

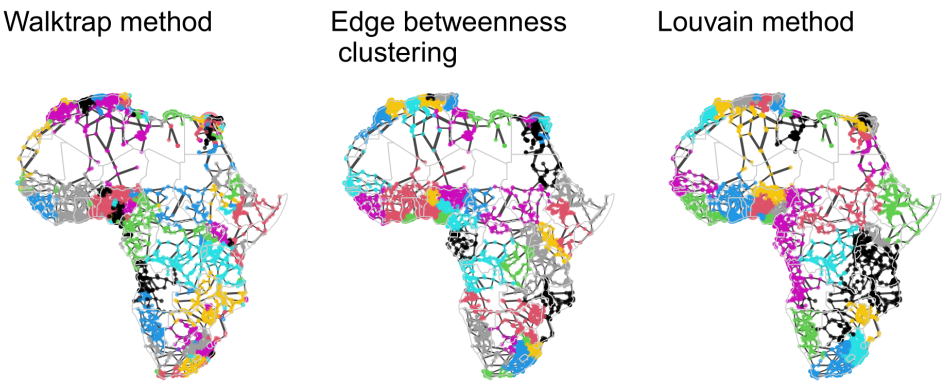


Figure 7: Outcomes of three community detection algorithms

```
  scale_size_continuous(range = c(.3, 4)) +  # Scale node size
  scale_color_identity() +  # Scale node color
  theme(legend.position = "none",
        panel.background = element_rect(fill = NA, colour = NA)) +
  geom_sf(data = world_subset,
          fill = NA,
          color = "gray80") +
  ggtitle("Louvain method")
```

In Figure 7, we visualise the results of the three community detection methods, i.e. walktrap, edge betweenness and the Louvain method. The results show that nodes that are geographically close and are part of densely connected clusters, generally belong to the same community, regardless of the method used. There is some degree of correspondance between the detected communities and the African countries.

[43]:
```
# Combine the three community detection algorithm plots
plot_wt + plot_eb + plot_mlc
```

[43]:
```
Output in Figure 7.
```

5.5 *Analysing network robustness*

Robustness is the ability of a network to maintain its basic functions in the presence of node and link failures (Barabási, Pósfai 2016). Percolation theory, originally developed in statistical physics to explore the relationship between microscopic and macroscopic properties of a medium, is a widely used approach to studying network robustness. It helps identify the conditions under which a network remains well-connected, ensuring that most nodes can still interact or communicate effectively despite potential disruptions. A well-connected network is characterised by the presence of a giant connected component (GCC), which is a single, large cluster of interconnected nodes (Bollobás 2001). The subset of network nodes within the GCC are reachable from each other, either directly or indirectly, through paths. This means that, from any node node within the GCC, it is possible to navigate through the network and eventually reach most of the other nodes.

The Molloy-Reed criterion is an important theoretical condition for the emergence of the GCC in a network (Molloy, Reed 1995). This criterion provides a threshold based on the degree distribution of nodes for a GCC to exist in a network. Specifically, the Molloy-Reed criterion states, in absence of degree correlations (no degree assortativity), that a GCC will emerge if the following is true: $\langle k^2 \rangle - 2\langle k \rangle > 0$, where $\langle k \rangle$ is the average degree and $\langle k^2 \rangle$ is the mean squared degree. The network is sufficiently connected to support the formation of a GCC when the inequality holds. Conversely, the network will fragment into small, isolated clusters when the inequality does not hold.

It is crucial in real-world applications to ensure that networks have a GCC for maintaining a system's resilience. For example, the GCC ensures that most locations remain accessible in transportation networks, while it enables the majority of devices or users to exchange information in communication networks. Understanding the robustness of these types of networks can therefore be valuable for improving regional resilience, because it allows regions to better accommodate shocks and develop new growth paths (Elekes et al. 2024).

More specifically, percolation analysis can be used to determine the threshold at which a GCC emerges. For instance, percolation analysis can help calculate the minimum number of edges that need to be added to keep the network in a well-connected state given a network with a specific number of nodes. On the other hand, inverse percolation (Barabási, Pósfai 2016) helps identify the point at which the removal of edges or nodes leads to network fragmentation, breaking it into smaller, isolated clusters. The following section focuses on inverse percolation, as the selected approach seeks to better understand network robustness by assessing a network's ability to remain well-connected when nodes or edges are removed.

A full inverse percolation algorithm or process is typically run so that the value of a percolation parameter that controls the removal of nodes or edges is updated in each iteration, and nodes or edges are removed accordingly. Key robustness metrics are measured in each iteration. One of the most used robustness metrics is the number of nodes in the GCC after the removal of nodes or edges. This metric is known as the size of the GCC. In many cases, we observe that abrupt changes occur in the size of the GCC for certain values of the percolation parameter. This indicates that some sort of failure occurs in the network that qualitatively changes its connectivity structure.

This type of analysis is demonstrated below. The percolation parameter of choice is the time of travel through each edge, accounting for the presence of borders. This variable is encoded by the `timeUCB` field in the `df_edges` data frame. Edges with `timeUCB` above the value of the percolation parameter are removed from the network in each iteration of the inverse percolation process. Furthermore, instead of considering the whole African road network, a subset is considered, as this facilitates timely execution of the algorithm. A subset formed by nodes and edges from the South region is used.

[44]:
```r
# Subset df_nodes for rows where Region is "South"
df_nodes_sub <- subset(df_nodes, Region == "South")

# Subset df_edges for rows where 'from' values are in Agglomeration_ID
df_edges_sub <- subset(df_edges,
                       from %in% df_nodes_sub$Agglomeration_ID)

# Subset df_edges_sub for rows where 'to' values are in Agglomeration_ID
df_edges_sub <- subset(df_edges_sub,
                       to %in% df_nodes_sub$Agglomeration_ID)
```

We can create an undirected graph from the redefined data frames of nodes and edges.

[45]:
```r
# Create an igraph graph 'g_sub' from 'df_edges_sub' and 'df_nodes_sub'
g_sub <- graph_from_data_frame(d = df_edges_sub,
                               vertices = df_nodes_sub,
                               directed = FALSE)
```

We can also visualise this subnetwork by running the code below. We will add the outlines of the countries in the South region as a base layer for this plot to give more geographical context to the above visualization. These are Botswana, Eswatini, Lesotho, Namibia and South Africa.

[46]:
```r
# Define a vector of target countries
south_countries <- c("Botswana", "eSwatini", "Lesotho",
                     "Namibia", "South Africa")
```

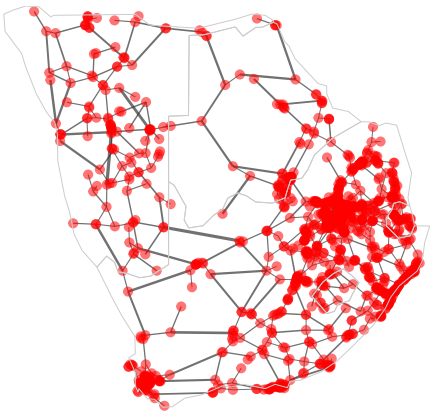


Figure 8: Visualisation of a road network in the South region of the African continent

```
# Subset the world map data to include only the target countries
world_south <- world[world$SOVEREIGNT %in% south_countries, ]

# Specify node layout
custom_layout <- data.frame(
  name = df_nodes_sub$agglosName,  # Node names from the graph
  x = df_nodes_sub$x,              # Custom x-coordinates
  y = df_nodes_sub$y               # Custom y-coordinates
)
```

[47]:
```
# Create plot
ggraph(as_tbl_graph(g_sub),
                    custom_layout) +  # basic graph plot
  geom_edge_link(color = "gray20",
                 alpha = 0.7,
                 aes(width = E(g_sub)$l * 0.1)) +  # custom edges
  scale_edge_width(range = c(.1, 0.7)) +  # scale edge size
  geom_node_point(aes(color = "red",
                      size = V(g_sub)$size,
                      alpha = 80)) +  # custom nodes
  scale_size_continuous(range = c(.1, 6)) +  # scale node size
  scale_color_identity() +  # scale node color
  theme(legend.position = "none",
        panel.background = element_rect(fill = NA, colour = NA)) +
  geom_sf(data = world_south,
          fill = NA,
          color = "gray80")  # basic map plot
```

[47]:
```
Output in Figure 8.
```

5.5.1 The inverse percolation algorithm

The inverse percolation algorithm simulates the progressive breakdown of a network by iteratively removing edges based on a percolation parameter (in this case, `timeUCB`). The algorithm evaluates how the removal of edges affects the structure of the network at each iteration. This process helps us understand how robust the network is and whether it disintegrates as we vary the percolation threshold. We need to store key metrics about the network at each step to analyse the results of this iterative process. We begin by creating empty data structures to store information about the network during each iteration. Specifically, we create four empty lists where we will record:

- The value of the percolation parameter at the current iteration.
- The size of the largest connected component (GCC) in the network, which indicates how much of the network remains connected.
- The total number of connected components in the network, which reflects how fragmented the network becomes.
- The average time to travel between any pair of nodes, weighted by the edge attribute `timeUCB`, to capture changes in the network connectivity.

These lists will each contain as many elements as iterations by the end of the inverse percolation process, corresponding to the results for all iterations.

[48]:
```
# Create empty vectors to store parameters, gccs, ncs, and times
parameters <- c()
gccs <- c()
ncs <- c()
times <- c()
```

We are now ready to perform the inverse percolation algorithm. The algorithm proceeds iteratively, with each iteration representing a step where edges with `timeUCB` values above the current percolation threshold are removed from the network. The logic for each step in the algorithm is as follows:

1) Define the current percolation threshold. At the start of each iteration, the current value of the percolation parameter `i` is defined. This value determines which edges will remain in the network at this step (those with `timeUCB` less than `i`).
2) Filter the network based on the percolation threshold. The nodes remain unchanged, but the edges are filtered to include only those with `timeUCB` values below the current threshold. This results in a modified edge list, which represents the network at the current percolation step.
3) Construct a new graph. Using the filtered edge list and the full node list, we create a new graph `g_perco`. This graph reflects the state of the network after removing edges based on the percolation parameter.
4) Analyse the modified graph by computing and storing the metrics defined before, i.e. the he size of the largest connected component (GCC), the number of connected components and the mean geographic road distance between nodes. Each metric is appended to its respective list for later analysis. By the end of the algorithm, these lists will contain a complete record of how the network evolved at each percolation step.

Below is the code implementation of the inverse percolation algorithm. Each line has been commented to describe its function:

[49]:
```
# Iterate over parameters
for (i in seq(0, max(df_edges_sub$timeUCB))) {

  # Create modified data frames based on the current parameter
  df_nodes_perco <- df_nodes_sub
  df_edges_perco <- subset(df_edges_sub, timeUCB < i)

  # Create a graph g_perco from the modified data frames
  g_perco <- graph_from_data_frame(d = df_edges_perco,
                                   vertices = df_nodes_perco,
                                   directed = FALSE)

  # Get connected components of the modified graph g_perco
  connected_components <- components(g_perco)

  # Append the current parameter value to the 'parameters' list
  parameters <- c(parameters, i)

  # Append the maximum connected component size to the 'gccs' list
```

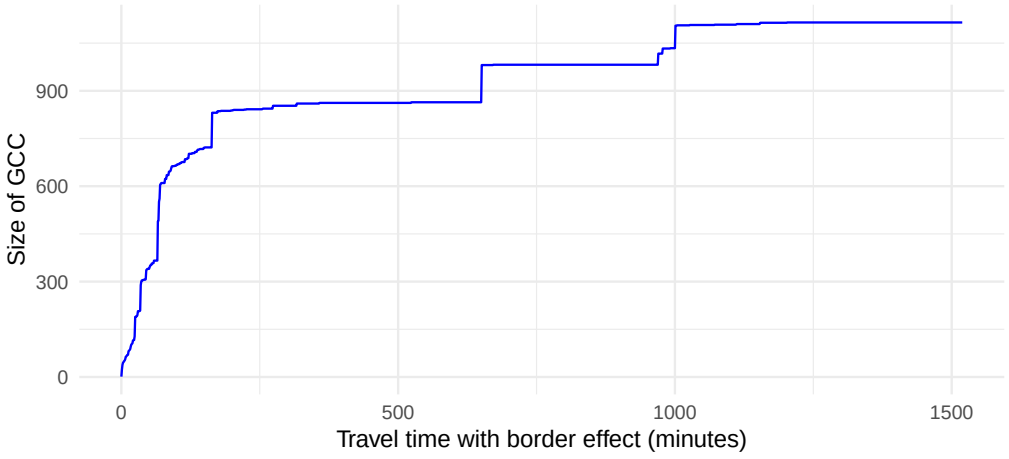


Figure 9: Relationship between the size of the Giant Connected Component and the percolation parameter

```
  gccs <- c(gccs, max(connected_components$csize))

  # Append the number of connected components to the 'ncs' list
  ncs <- c(ncs, connected_components$no)

  # Calculate and append the mean distance weighted by timeUCB
  times <- c(times, mean_distance(g_perco,
                                  directed = FALSE,
                                  weights = df_edges_perco$timeUCB,
                                  unconnected = TRUE))
}
```

5.5.2 Changes in the size of the giant connected component as edges are removed

Once the algorithm is done running, we can plot the size of the GCC as the value of the percolation parameter is varied.

[50]:
```
# Create a data frame for the plot with parameters and gccs
df <- data.frame(x = parameters, y = gccs)
# Create a ggplot2 plot with customised aesthetics and labels
ggplot(data = df, aes(x = x, y = y)) +
  geom_line(color = "blue") +
  labs(x = "Travel time with border effect (minutes)",
       y = "Size of GCC") +
  theme_minimal()
```

[50]:
```
Output in Figure 9.
```

We observe for small values of the percolation parameter that rapid changes occur in the size of the GCC. There are sudden changes in the size of the GCC when the percolation parameter takes approximately the values 150, 650, 1000, indicating that there has been a significant alteration in the network's topology. For example, when edges with associated travel times of 1000 minutes or less are removed, nodes that act like hubs may lose connections, and the size of the GCC becomes smaller as a result.

5.5.3 Changes in the number of connected components as edges are removed

We can also plot the number of connected components as the value of the percolation parameter is varied.

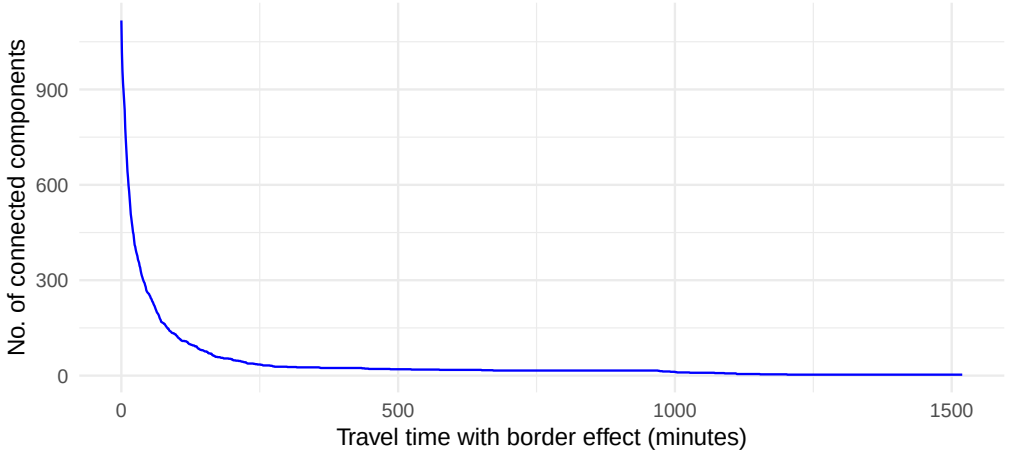


Figure 10: Relationship between the number of connected components and the percolation parameter

[51]:
```
# Create a data frame for the plot with parameters and ncs
df <- data.frame(x = parameters, y = ncs)

# Create a ggplot2 plot
ggplot(data = df, aes(x = x, y = y)) +
  geom_line(color = "blue") +
  labs(x = "Travel time with border effect (minutes)",
       y = "No. of connected components") +
  theme_minimal()
```

[51]:
```
Output in Figure 10.
```

We observe that nearly all the edges in the network are removed for small values of the percolation parameter so that there are as many components as there are nodes. We also see that the number of connected components is reduced if the percolation parameter is above 250 minutes, highlighting that the connectivity of the network is greater above that threshold value as the network is less fragmented.

5.5.4 Changes in the average travel time as edges are removed

Finally, we plot the average travel time between any pair of nodes as the value of the percolation parameter is varied.

[52]:
```
# Create a data frame for the plot with percolation parameter and times
df <- data.frame(x = parameters, y = times)

# Create a ggplot2 plot
ggplot(data = df,
       aes(x = x, y = y)) +
  geom_line(color = "blue") +
  labs(x = "Travel time with border effect (minutes)",
       y = "Average travel time (minutes)") +
  theme_minimal()
```

[52]:
```
Output in Figure 11.
```

Note that all the edges are removed when the percolation parameter is 0, and so the corresponding value of the average travel time is NA. As the percolation parameter is increased, less edges are removed from the original network and more possible paths

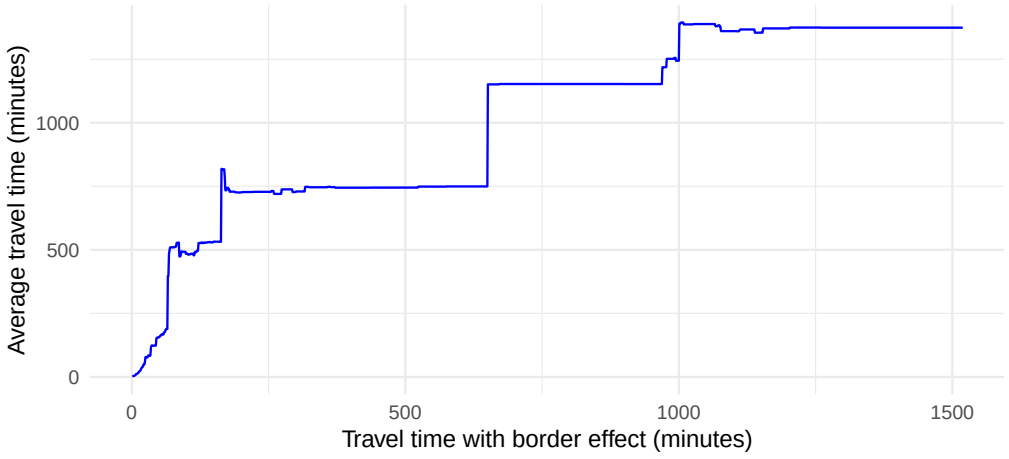


Figure 11: Relationship between the average travel time with border effect (minutes) and the percolation parameter.

arise. Note that the average travel time is only computed for existing paths (hence the `unconnected=TRUE` parameter in the `mean_distance()` function). The sudden changes in the average travel time and the sudden changes in the size of the GCC happen for the same values of the percolation parameter. For example, there is a large increase in the average travel time as the percolation parameter decreases below approximately 650 minutes. This suggests that the two parts of the network that were connected for higher values of the percolation parameter become unconnected for values below 650 minutes as a considerable number of edges get removed. As a result, the average travel time increases since there are possibilities to travel further.

6 Conclusion

This article provides an introduction of key concepts, definitions, and applications of network analysis for the study of geographic data, as well as instructions for the practical implementation of network analysis in R. The approach presented here offers deeper insights into how location, distance, and spatial distribution impact the behaviour of systems where connections are important by integrating geographic data with network analysis methods.

The versatility of these methods highlights their potential for application across diverse geographic contexts. While the African road network serves as a specific example, the underlying principles can be adapted to different contexts and types of networks. This adaptability makes it possible to apply network analysis to various scales, from small scale street networks (Ma et al. 2024), to intraurban transport networks (Zhong et al. 2014), inter-regional (Arcaute et al. 2016), or global scale spatial networks (Colizza et al. 2006).

The methods presented here will serve as a foundational resource for the study of geographic data as the field of spatial network analysis continues to evolve. The integration of theoretical concepts with practical, accessible tools ensures that researchers and practitioners can apply these techniques to address the current challenges. Ultimately, this work contributes to the ongoing development of more robust, efficient, and sustainable networks in an increasingly interconnected world.

References

Aldous JM, Wilson RJ (2000) *Graphs and Applications*. Springer London. CrossRef

Anselin L (1988) *Spatial Econometrics: Methods and Models*. Springer Netherlands. CrossRef

Arcaute E, Molinero C, Hatna E, Murcio R, Vargas-Ruiz C, Masucci AP, Batty M (2016) Cities and regions in Britain through hierarchical percolation. *Royal Society Open Science* 3: 150691. CrossRef

Barabási A, Pósfai M (2016) *Network science*. Cambridge University Press. http://barabasi.com/networksciencebook/

Barthelemy M (2018) *Morphogenesis of Spatial Networks*. Lecture Notes in Morphogenesis. Springer International Publishing. CrossRef

Barthelemy M, Boeing G (2024) A review of the structure of street networks. *Findings*. CrossRef

Barthélemy M (2011) Spatial networks. *Physics Reports* 499: 1–101. CrossRef

Batty M (2007) *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. The MIT Press

Bettencourt L (2021) *Introduction to Urban Science*. The MIT Press. CrossRef

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008. CrossRef

Bollobás B (2001) The evolution of random graphs – The giant component. In: Bollobás B (ed), *Random Graphs*. Cambridge University Press, 130–159. CrossRef

Cabrera-Arnau C, Zhong C, Batty M, Silva R, Kang SM (2023) Inferring urban polycentricity from the variability in human mobility patterns. *Scientific Reports* 13. CrossRef

Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences* 103: 2015–2020. CrossRef

Curtin KM (2007) Network analysis in geographic information science: Review, assessment, and projections. *Cartography and Geographic Information Science* 34: 103–111. CrossRef

Elekes Z, Tóth G, Eriksson R (2024) Regional resilience and the network structure of inter-industry labour flows. *Regional Studies* 58: 2307–2321. CrossRef

Flake GW (1998) *The computational beauty of nature*. MIT Press, Cambridge, MA, USA

Fortunato S, Newman MEJ (2022) 20 years of network community detection. *Nature Physics* 18: 848–850. CrossRef

Franklin RS (2020) Geographical analysis at midlife. *Geographical Analysis* 53: 47–60. CrossRef

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99: 7821–7826. CrossRef

Guimerà R, Amaral LAN (2004) Modeling the world-wide airport network. *The European Physical Journal B - Condensed Matter* 38: 381–385. CrossRef

Hartshorne R (1939) The nature of geography: A critical survey of current thought in the light of the past. *Annals of the Association of American Geographers* 29: 173. CrossRef

Kolaczyk ED, Csárdi G (2014) *Statistical Analysis of Network Data with R*. Use R! Springer New York. CrossRef

Lawrence J, Blackett P, Cradock-Henry NA (2020) Cascading climate change impacts and implications. *Climate Risk Management* 29: 100234. CrossRef

Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ (2012) Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS ONE* 7: e31743. CrossRef

Ma D, He F, Yue Y, Guo R, Zhao T, Wang M (2024) Graph convolutional networks for street network analysis with a case study of urban polycentricity in Chinese cities. *International Journal of Geographical Information Science* 38: 931–955. CrossRef

Maier G (2014) OpenStreetMap, the Wikipedia map. *REGION* 1: R3–R10. CrossRef

Miller HJ (2017) Geographic information science II: Mesogeography. *Progress in Human Geography* 42: 600–609. CrossRef

Miller HJ, Clifton K, Akar G, Tufte K, Gopalakrishnan S, MacArthur J, Irwin E (2021) Urban sustainability observatories: Leveraging urban experimentation for sustainability science and policy. *Harvard Data Science Review*. CrossRef

Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* 6: 161–180. CrossRef

Moriconi-Ebrard F, Harre D, Heinrigs P (2016) *Urbanisation Dynamics in West Africa 1950–2010*. CrossRef

Nelson T, Frazier AE, Kedron P, Dodge S, Zhao B, Goodchild M, Murray A, Battersby S, Bennett L, Blanford JI, Cabrera-Arnau C, Claramunt C, Franklin R, Holler J, Koylu C, Lee A, Manson S, McKenzie G, Miller H, Oshan T, Rey S, Rowe F, Şalap Ayça S, Shook E, Spielman S, Xu W, and JW (2025) A research agenda for GIScience in a time of disruptions. *International Journal of Geographical Information Science* 39: 1–24. CrossRef

OpenStreetMap Contributors (2023) Openstreetmap. https://www.openstreetmap.org/ [accessed 30-08-2023]

O'Sullivan D, Manson SM (2015) Do physicists have geography envy? And what can geographers learn from it? *Annals of the Association of American Geographers* 105: 704–722. CrossRef

Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: Yolum p, Güngör T, Gürgen F, Özturan C (eds), *Computer and Information Sciences - ISCIS 2005*. Springer, Berlin, Heidelberg, 284–293. CrossRef

Prieto Curiel R, Cabrera-Arnau C, Bishop SR (2022) Scaling beyond cities. *Frontiers in Physics* 10. CrossRef

Prieto-Curiel R, Heo I, Schumann A, Heinrigs P (2022) Constructing a simplified interurban road network based on crowdsourced geodata. *MethodsX* 9: 101845. CrossRef

Rand WM (1971, December) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66: 846–850. CrossRef

Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. *Policy Sciences* 4: 155–169. CrossRef

Strano E, Giometto A, Shai S, Bertuzzo E, Mucha PJ, Rinaldo A (2017) The scaling structure of the global road network. *Royal Society Open Science* 4: 170590. CrossRef

Uitermark J, van Meeteren M (2021) Geographical network analysis. *Tijdschrift voor Economische en Sociale Geografie* 112: 337–350. CrossRef

Zhong C, Arisona S, Huang X, Batty M, Schmitt G (2014) Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science* 28: 2178–2199. CrossRef

REGION
The Journal of ERSA
Powered by WU

Volume 12, Number 1, 2025, 103–116
DOI: 10.18335/region.v12i1.540

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

Identifying and analyzing logistics land use: a case study of the Rhineland Metropolitan Region

Andre Thiemermann[1], Florian Groß[1]

[1] University of Wuppertal, Wuppertal, Germany

Received: 14 May 2024/Accepted: 13 May 2025

Abstract. In Germany, the identification of logistics land is rarely done, among other things due to the anonymization of employment and building data. The paper at hand gives an overview of data sources used for the identification of spatial patterns of logistics facilities and presents a method for identifying logistics land based on publicly available data, to present an image of the existing spatial structure of logistics land. Identified spatial hotspots are mostly located in Metropolises/Regiopolises and their suburbs, along highways in areas with flat relief, and in the vicinity of large inland terminals/inland harbors.

1 Introduction

Logistics activities account for a large part of land consumption. In Germany, warehouse buildings contributed about 25% of all built floor space of non-residential buildings in 2018 (Kretzschmar et al. 2021). Construction activity concentrates on a few municipalities with good transport infrastructure connections (ibid.).

Concerning the investigation of spatial patterns of logistics facilities, extensive international studies are available. Nevertheless, only Holl, Mariotti (2017) provide an overview of the methods and databases used for individual international studies, which, however, does not include German-language studies and does not focus in detail on the databases used. In Germany, there still exist only aggregated studies of logistics real estate at the NUTS 3 level (corresponding to the level of counties) or municipality level (e.g. Klauenberg, Krause Cauduro 2020, Busch 2013, Kretzschmar et al. 2021), which is also due to problems with the availability and quality of relevant data sources. In particular, small-scale studies at the level of individual logistics facilities and their respective estates have not yet been carried out.

We want to close this gap in the following article. We present a dataset that, for the first time, shows small-scale spatial patterns of logistics estates in a German study area. Based on the developed dataset, we exemplarily identify hotspots in logistics land[1].

The paper is structured as follows. Section 2 introduces the background and thereby provides inter alia an overview of data sources used for the identification and examination of spatial patterns of logistics facilities. Section 3 introduces the study area. After that, Section 4 presents the methods of identifying logistics land and further data preparation. Section 5 presents the examination of spatial autocorrelation, inter alia, the identification of hotspots of logistics land. Section 6 draws conclusions based on these examinations.

[1]The dataset can be download from https://doi.org/10.57806/dkmd9hk5

2 Background

In this section, we provide an overview of existing databases that are used in the study of spatial patters of logistics facilities and the specific difficulties that exist in this regard for German study areas.

2.1 Existing data on spatial patterns of logistics facilities

The existing analyses of spatial patterns of logistics facilities can essentially be broken down into

- company/employment data
- data on construction activity/property trading
- survey data
- development of own databases/data fusion.

There are also special cases, such as the use of GPS tracks (see e.g. Trent, Joubert 2022). Apart from a few studies like Busch (2013), Jaller et al. (2022), Heitz et al. (2019), or Nefs et al. (2024), there is usually no consolidation of data.

2.1.1 Company/employment data

Company and employment enable a view of logistic-specific companies and employment in small-scale areas such as municipalities. The classification of economic activity is generally used to identify data on logistic-specific employment and companies. For that, the Europe-wide NACE classification can be used.

Jaller et al. (2022) base their analyses in Southern California on the US Zip Code Business Patterns (ZBP), which provide information on the number of companies and employment (here NAICS 493: Warehousing and Storage; comparable to NACE 52.1) between 1998 and 2016 at the postcode district level. Holguín-Veras et al. (2022) uses the same data basis but also uses a categorization of all economic sectors into freight-intensive and service-intensive sectors from Holguín-Veras et al. (2016).

Klauenberg, Krause Cauduro (2020) and Heitz et al. (2017) base their analyses in Berlin/Brandenburg and the metropolitan regions of Paris and Randstad on aggregated company figures for NACE category 52.1 (warehousing) at the municipal level. Heitz, Beziat (2016) supplement their data from the business register for parcel services (NACE categories 52.29 and 53.20) with additional interviews with parcel service providers, as it is clear that only some of the locations of parcel service providers can be identified from the data.

The employment data is also classified in the International Standard Classification of Occupations (ISCO-08). Busch (2013) uses this classification and queries the number of employees according to various NACE aggregates (including retail (NACE G), transport and storage (NACE 49-52), CEP (NACE 53)) for the occupational group 513 'Warehousing, postal and delivery services and goods handling' for German NUTS 3 areas. Due to anonymization regulations, such surveys in Germany can only be carried out at this aggregated spatial level.

2.1.2 Survey data

In their study, Sakai et al. (2015) use a freight survey in the Tokyo metropolitan region from 2003, which includes 4,109 responses (including 2,803 responses with >400 m² of floor space) from companies that use logistics facilities. The advantage here is that, in addition to data on the respective logistics facilities, logistical behavioral data can also be queried. Thus real origin-destination-relations can also be depicted. The following data on the respective companies were collected as part of the survey:

- year of construction of the logistics facility used
- tonnes transported
- truck trips generated
- goods handled
- origin and destination of the shipments.

2.1.3 Data on construction activity/real estate industrial transactions

Another option, particularly for the observation of time series, is the use of transaction data in property trading or the observation of construction activity of logistics facilities.

Jaller et al. (2022) do the former: they look at transaction data for properties in Southern California between 1989 and 2018. Inter alia, these are typified by warehouse buildings and distribution buildings. Busch (2013) and Kretzschmar et al. (2021) analyze the construction activity of logistics facilities in Germany by looking at completed warehouse buildings in the statistics on construction work completed, which are part of the construction activity statistics. The main challenge here is that the identification of warehouse buildings differs depending on the relevant state statistics authority (Busch 2013). Furthermore, the reported time of building completion often does not correspond to the actual time; instead, there are often delays in reporting (Kretzschmar et al. 2021), which leads to inaccuracies.

2.1.4 Development of own databases/data fusion

Heitz et al. (2019) and Nefs et al. (2024) generate their own databases by fusing datasets.

Heitz et al. (2019) implement this for the Paris metropolitan region and justify their approach by arguing, among other things, that no clear allocation to the respective logistics segment can be derived solely from the allocation to the economic activity. They use a dataset comparable to the German business register and a list of large French warehouses (Répertoire des Entrepôts) as a basis. Specific buildings are validated with aerial and street images. In addition, areas, where logistics land uses are to be expected, are searched manually, and aerial photographs and planning documents are scrutinized. The identified buildings are geocoded, and further information is added. This includes

- function of the logistics facility under consideration
- type of logistics company that operates the logistics facility under consideration
- goods processed/transhipped there (specific (beverages, food, equipment), generic (e.g., parcels, general cargo))
- destinations of the processed/transhipped goods (households, companies by sector)

Nefs et al. (2024) implement this for the Netherlands and generate a time series of logistics facilities and their respective estate for the period from 1980 to 2021. Microdata from the official Dutch business register is used for this purpose, which provides information on specific company locations, the number of employees, and the specific NACE classification of economic activity. In addition, another source on current and planned commercial estates and OpenStreetMap is used. The year of construction is also obtained from a building administration dataset. Furthermore, a visual validation is carried out, in particular, to take into account newly built, very large distribution centers, based on Google Streetview.

To extrapolate the land use of warehouse buildings at the level of NUTS 3 areas, Kretzschmar et al. (2021) use, among other things, information on the floor space of completed buildings from the construction activity statistics, and - to calculate standard values for the ratio between building footprint and total estate occupied - building footprints from an official building dataset and land lots used industrially or commercially from the Cadastre Information System ALKIS.

2.1.5 Special case: GPS tracks of truck trips

One in literature discussed effect of the phenomenon of logistics sprawl is the increasing number of truck mileage. Data on real trips to/from logistics facilities can be used to empirically analyze this much-discussed effect when viewed over time. To date, there have been very few studies on this. Trent, Joubert (2022) use GPS tracks of around 16,000 vehicles used for commercial transport (1-2% of the total fleet) in South Africa between 2010 and 2014, which depict journeys in the metropolitan regions of Gauteng, Cape Town, and eThekwini.

2.1.6 Additional secondary data

Further data are used when analyzing spatial patterns of logistics facilities. These essentially include:

- socio-demographic data (included in Jaller et al. 2022, e.g. population, median age, proportion of white population, median household income, median household value, public transport users)
- Data on infrastructure relevant to freight transport (locations of CT terminals in Jaller et al. 2022)
- Land use (designated commercial areas in Sakai et al. 2020)

Jaller et al. (2022) also draw on an environmental index at the level of US postcode districts, which combines indices on environmental pollution (exposure and environmental effects) and population characteristics (including socio-economic factors) at the level of ZIP code districts in California.

2.2 Difficulties in the study of spatial patterns of logistics in Germany

In Germany, there are currently few studies analysing spatial patterns of logistics, e.g. Klauenberg, Krause Cauduro (2020). This is (so far) mainly due to several problems that arise from a lack of data availability:

- extensive anonymization of employment data on municipality level,
- according to Busch (2013), no clear categorization of a given building concerning the economic activity in building statistics, different from the Dutch and French dataset examined by Heitz et al. (2017),
- no georeferenced data on logistics buildings over multiple periods, such as those used e.g. by Dablanc, Rakotonarivo (2010).

3 Study area

The Rhineland metropolitan region is located in the West of the German state of North Rhine-Westphalia (NRW) and includes roughly the two administrative governmental districts of Cologne and Düsseldorf. It includes the metropolises of Dusseldorf and Cologne and the Western part of the Ruhr area. The metropolitan region has a share of 36% of the area of North Rhine-Westphalia and, with a population of about 9 million inhabitants, a share of 50% of the inhabitants. Relevant population growth is particularly evident in the metropolitan areas and their suburbs. Unlike many agglomerations of its size, the Rhineland metropolitan region has a polycentric spatial structure. The metropolitan cores here are Dusseldorf and Cologne.

As a central agglomeration in Central Europe, the Rhineland Metropolitan Region is the location of important transport hubs and at the same time an internationally significant business location and sales market. The metropolitan region is both a relevant source and destination for freight traffic, as well as transit traffic. Together, they contribute to a high utilization and sectional congestion of the road and rail networks and thus to a high demand for logistics space for transshipment and warehousing.

A survey among logistics stakeholders of the Rhineland Metropolitan region, as included in Leerkamp et al. (2022), has revealed the biggest weaknesses of the region in terms of logistics locations (see Figure 1). The most mentioned weakness is congestion of the road network (ibid.). Aspects that contribute to logistics sprawl, like insufficient land availability, high land prices, and lack of support from the public sector are also of high relevance. In this context, the dynamic development of e-commerce will also continue to drive the demand for locations for distribution centers as well as parcel sorting centers. Accordingly, logistics companies expect a high demand for locations on the outskirts and increasingly in the core areas of the region's major cities, too (see Figure 2).

4 Identifying logistics land and data preparation

The identification of logistics facilities and estates occupied by them is essentially based on the identification of estates on which logistics buildings are located. This procedure is

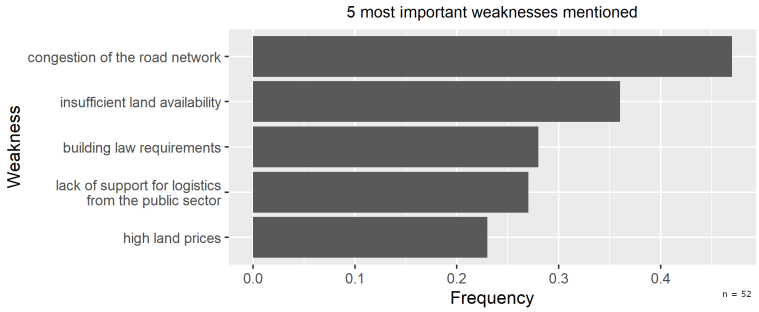


Figure 1: Most mentioned weaknesses of the Rhineland Metropolitan region as a logistics region (Leerkamp et al. 2022)

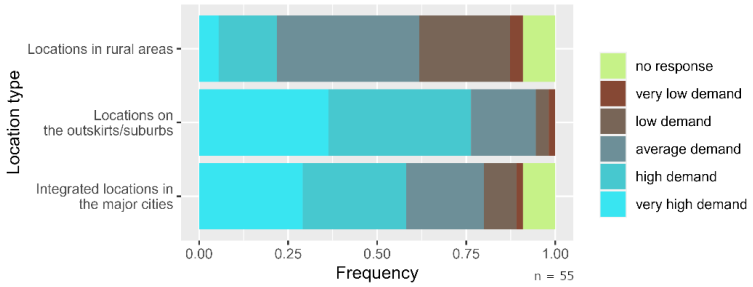


Figure 2: Demand location types for logistics facilities in the study area in the upcoming 5-10 years (Leerkamp et al. 2022)

based on Kretzschmar et al. (2021). In contrast to Kretzschmar et al. (2021), who use logistics facilities identified in exemplary NUTS 3-areas to calculate standard values for the ratio between building footprint and total estate occupied, here the individual logistics facilities identified are themselves analyzed on a small scale (1 km² grid) for an entire metropolitan region. The fact that several data sources are used with the cadastral data and OpenStreetMap means that the procedure can also be described as data fusion. In the procedure, datasets of estates are merged with datasets of logistics buildings. Figure 3 summarizes the procedure explained in the following.

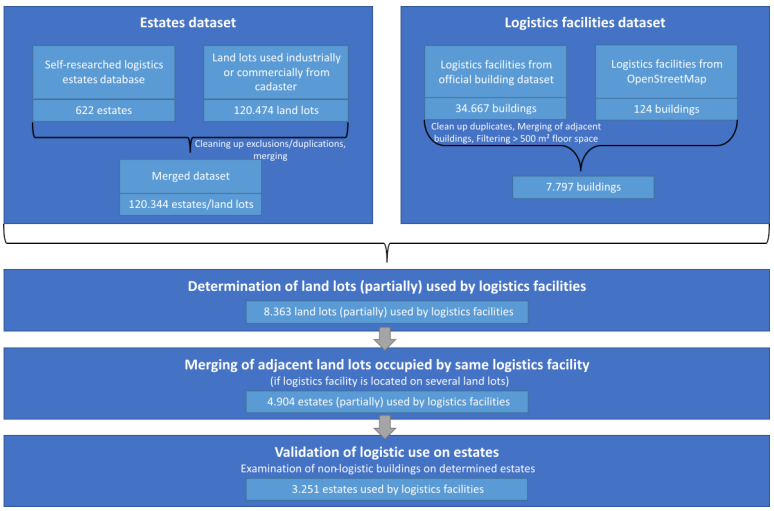


Figure 3: Identification of logistics estates

4.1 Used datasets and their preparation

Regarding the estates, two datasets are used. On the one hand, an already existing self-researched logistics estate database created by Leerkamp et al. (2022) and contains 622 existing logistics estates that are bigger than 2 ha. As a basis for the identification of further logistics estates, land lots used industrially or commercially are utilized from the official Cadastre Information System ALKIS (source: Geobasis NRW 2020). After merging and cleaning up the two datasets, 120,344 land lots/estates remain.

The logistics buildings are extracted from two datasets. The first one is an official building dataset (source: Geobasis NRW 2021), that is freely accessible in the state of North Rhine-Westphalia. This dataset also contains detailed information on the function of each building, so logistics buildings can be identified. However, detailed consideration shows that the categorization of the building function is not consistent throughout the region, so some logistics facilities cannot be identified by the official designation of the building function. Therefore, further logistics buildings were extracted from OpenStreetMap (source: OSM 2021).

The first dataset is edited to this effect, that duplicates are removed, adjacent buildings are merged, and very small buildings (< 500 m² floor space, proceeding according to Busch 2013) are removed. After merging the two datasets, 7,797 logistics buildings remain.

4.2 Identification of logistics land

The further identification of logistics land consists of three procedural steps.

The first step is the determination of land lots that are fully or partially occupied by logistics facilities (see left part of Figure 4). As a result, 8,363 land lots remain.

The second step is the generation of estates from the land lots that are (partially) occupied by logistics facilities (see top right in Figure 4). Hence, adjacent land lots, that are occupied by the same logistics facility are merged into one estate. Consequently, 4,904 estates remain, that are at least partially occupied by a logistics facility.

In the third and final step, logistics use of the estates under consideration and that have not been identified by own research is validated. For this purpose, the share of building floor space, that is used by logistics facilities is calculated for each estate (see bottom right in Figure 4). If the share is below 75%, the estate under consideration is removed. This is to ensure that only estates for which logistics is the primary function are considered.

As a result, 3,251 logistics estates are obtained and remain for further examination. Figure 5 shows the size distribution for the estates itself and the building footprints. The median of the estate size is 0.5 ha, whereas the median of the building footprint is 0.17 ha.

Figure 4: Procedural steps for the identification of logistics land

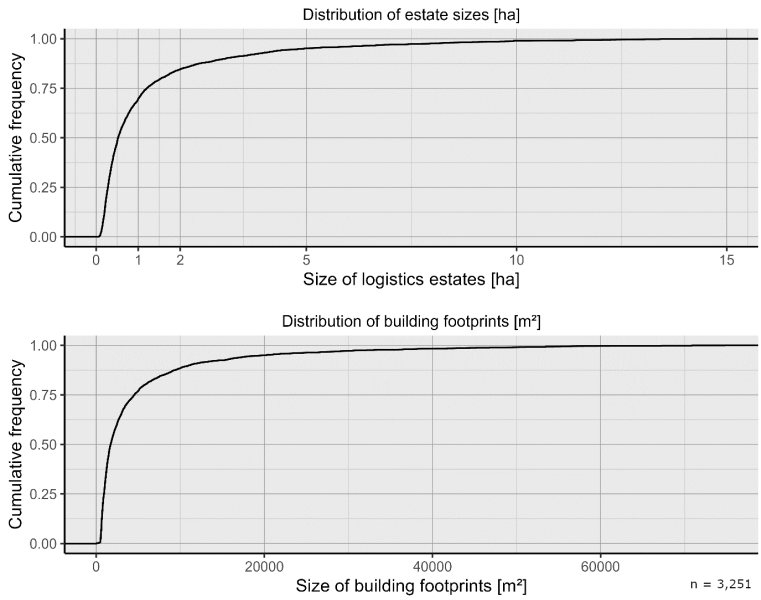


Figure 5: Size distribution of building footprints and estates of logistics land in the study area

4.3 Checking the dataset for completeness

As already mentioned, the categorization of the building function in the official dataset is not uniform across the board, meaning that not all existing logistics facilities can be identified from this dataset. This is probably due to the fact that the associated survey is the responsibility of the cadastre authorities, whose focus is not on the differentiated consideration of logistics.

Through the additional use of OpenStreetMap and the development of our own database, we were able to identify additional logistics space that would not have been identified if we had only used the official building dataset. Figure 6 shows an example of a logistics estate that could only be identified as a logistics estate through the use of OpenStreetMap. A total of 455 additional logistics estates were identified in this way; this corresponds to 14% of all identified logistics estates.

As there is no knowledge of the whole population of logistics buildings, it can be assumed that the dataset does not represent a complete picture of logistics land use in the area under investigation. Nevertheless, a qualitative comparison with known logistics facilities shows that this procedure represents an approach that can be used to at least roughly determine land use by logistics for this large study area.

4.4 Spatial aggregation and further variables

In the further, a 1 km² grid is used, because it also enables examinations regarding spatial autocorrelation. Therefore, further variables like the driving distance to the next inland terminals are adapted to this grid. The use of the raster also allows to consider further variables like the existing area of industrial/commercial land in each grid cell.[2] The variables used are presented in Table 1. It should be noted that employment figures were generally used at the county level due to the high level of anonymization at the municipal level.

In 1,623 of 12,773 grid cells logistics land can be determined by the method described above. Figure 7 shows the grid cells that contain identified logistics land. Evidently, they concentrate on the river Rhine, especially around the metropolises Cologne and Dusseldorf, the inland port of Duisburg, and flat suburban/exurban areas west of Dusseldorf and

[2]Standard land values for industrial/commercial land could not be determined for all grid cells with existing logistics land.

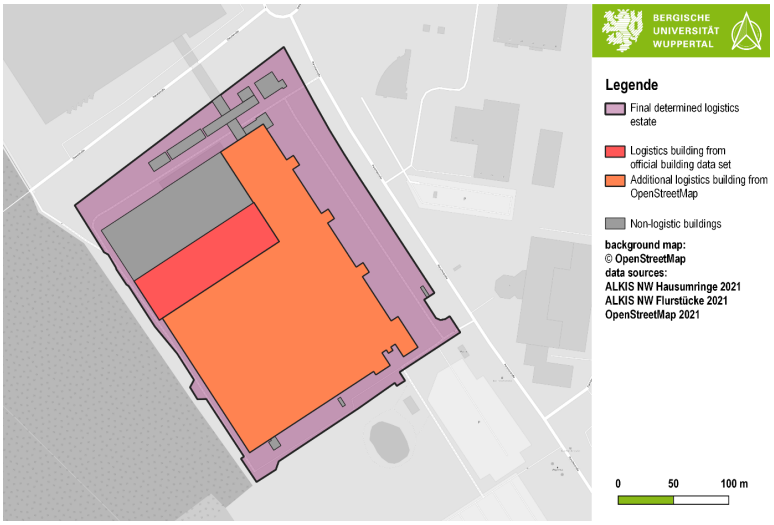


Figure 6: Example for gaps in the official building dataset

Cologne. As Figure 8 shows, these areas also account for a very large share of the total logistics land identified in the study area.

Table 1: Variables used for examination

Variable type	**Variable description**	**Abbrevation**	**Aggregation of variable/ starting point**	**Data source**
Demographics	Population density (inh./km²)	pop_dens	Population density on municipality level	BKG (2023)
Accessibility	Driving distance [km] to next high-order center	dist_hoc	Centroid of grid cell	Own calculation
	Driving distance [km] to next inland terminal	dist_terminal	Centroid of grid cell	Own calculation
	Driving distance [km] to next motorway/trunk access	dist_motorway	Centroid of grid cell	Own calculation
Employment/ Establish-ments	Employment in NACE-category 49.4, 52, 53 on county level in 2019	log_emp_county	County level (assignment of grid cell based on centroid)	IT.NRW (2021)
	Establishments in NACE-category 49.4. 52, 53 on municipal level in 2019	log_est_mun	Municipal level (assignment of grid cell based on centroid)	IT.NRW (2021)
	Employees occupied in warehousing, mail and delivery, cargo handling (all NACE-categories) on county level in 2020	wmc_emp_county	County level (assignment of grid cell based on centroid)	BA (2021)
Land market	Standard land value for all industrial/ commercial land in 2021	land_value_ind	Average for indus-trial/commercial land in grid cell	Own calculation based on Bezirksregierung Köln (2021)

Continued on next page

Table 1: Variables used for examination (continued)

Variable type	Variable description	Abbrevation	Aggregation of variable/starting point	Data source
	Existing area of industrial/commercial land in 2021	land_ind	Sum of industrial/commercial land area in grid cell	Own calculation based on Geobasis NRW (2020)
Logistics land	Area [ha] of logistics land identified in grid cell	log_land	Sum of grid cell	Own calculation

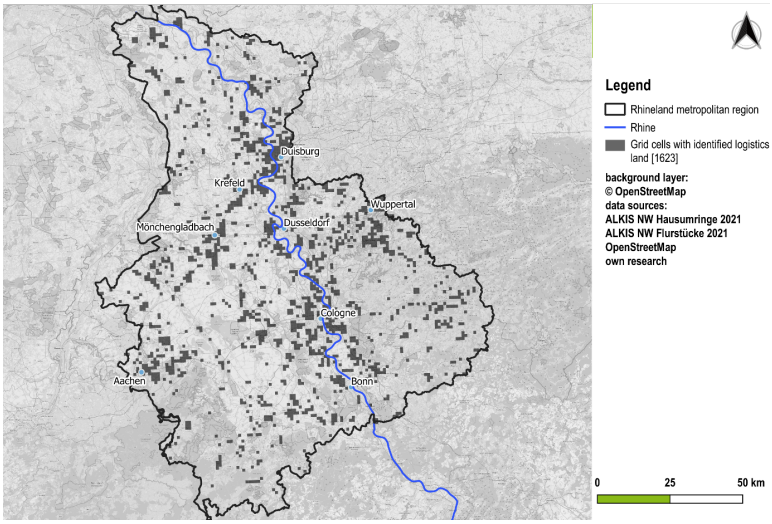


Figure 7: Grid cells with identified logistics land

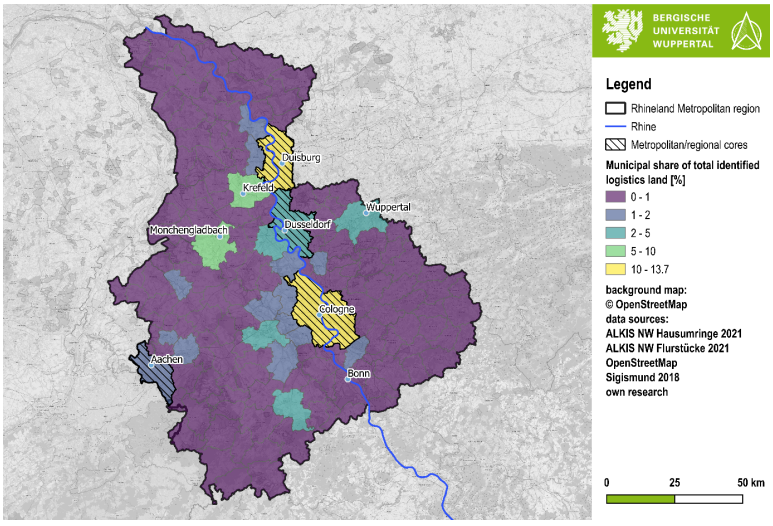


Figure 8: Municipal share of total identified logistics land

5 Analysis of spatial autocorrelation

As an example of the further usability of the dataset, measures of spatial autocorrelations are calculated in the following. Spatial autocorrelation is an application area of spatial statistics. According to Cliff, Ord (1970), spatial autocorrelation is defined as: "If the

presence of some quantity in a county (sampling unit) makes its presence in neighboring counties (sampling units) more or less likely, we say that the phenomenon exhibits spatial autocorrelation." Here, the global and local spatial autocorrelation is calculated for the appearance of logistics land.

5.1 Global spatial autocorrelation

For global spatial autocorrelation, the common Moran's I measure is used. It represents a weighted correlation, with weights increasing in spatial distance (Kirilenko 2022). It thereby measures autocorrelation over an entire area under consideration (ibid.). The Moran's I value can range from -1 to +1 (O'Sullivan, Unwin 2010, p. 206). Values greater than +0.3 indicate a strong positive spatial autocorrelation, whereas values less than -0.3 indicate a strongly negative spatial autocorrelation (ibid.).

Table 2 presents the results for the Global Moran's I-Index. With a statistically significant value of 0.235, the calculated Moran's I is close to a strong positive spatial autocorrelation (Moran's I $>$ 0.3). That means the spatial pattern of logistics land in the Rhineland metropolitan region is likely to be not random.

Compared with the results of Jaller et al. (2022), who calculated the Moran I for warehouses and distribution centers they got from the Zip Code Business Pattern database for five metropolitan regions in California in 2016, the index of the Rhineland metropolitan area is similar to Southern California (0.24) that contains e.g. Los Angeles and Orange County. The only metropolitan region that has a higher index, i.e. even more concentrated logistics facilities, is San Joaquin County (0.36).

Table 2: Results for global spatial autocorrelation of the Rhineland metropolitan region

Indicator	**Value**
Number of raster cells	12,773
Number of raster cells that contain logistics land	1,623
Global Moran's I-Index	0.235
Standard deviation	52.76
p-Value	$<$ 0.001

5.2 Local spatial autocorrelation

The hotspot analysis by Getis, Ord (1992) is in contrast a local measure, i.e. it is calculated for each object of investigation individually. This allows the determination of local concentrations of high or low values of an attribute (O'Sullivan, Unwin 2010). For each object i, the value G_i is calculated, which represents the share of the sum of all attribute values (e.g. logistics employment), which is represented by the neighbors of object i (located in a defined distance). Accordingly, G_i will be high for objects where high values accumulate (Getis, Ord 1992). With the slightly modified G_i^*, the values of the object i itself are also included in the consideration (ibid.). The main result is the z-score (corresponding to the z-transformation), which is the difference between the calculated G_i and the expectancy-value of G_i in the ratio to the standard deviation of the calculated G_i (ibid.). A high z-score value indicates that large values of the attribute under consideration are concentrated around the location under consideration (ibid.). This must additionally be tested for statistical significance (ibid.). Busch (2013) uses this, for example, to identify hot spots in the spatial distribution of logistics employment. In this case, the G_i^* is calculated contiguity-based, using the Queen's Case, i.e. for each grid all neighbors are considered, even those touched only at a single point. The respective area of logistics land in each grid cell is used as an attribute value.

As a result of the hot-spot analysis, 717 statistically significant hotspots can be identified, i.e. around these grid cells a high number of logistics land is concentrated. In some cases, grid cells with no logistics land are recognized as hot-spots, because they are adjacent to grid cells with high numbers of logistics land. These hotspots account for 65.4% of all logistics land identified in the study area. The 20 cells with the highest z-score, i.e. where logistics land use is highly concentrated and therefore logistics clusters

exist, are located in the area of the Port of Duisburg, in a logistics park in the proximity of an inland terminal in Duisburg and a logistics park in Monchengladbach that is purely geared towards trucking.

Altogether, it can be demonstrated that the hot-spots concentrate on four regions/facilities (see Figure 9):

- Metropolises/Regiopolises and their suburbs (e.g. Aachen, Cologne)
- Along highways in sub-/exurban areas with flat relief (west of Cologne/Dusseldorf and south of Monchengladbach)
- Large inland terminals/inland harbors (above all Duisburg).

Looking only at the grid cells containing identified logistics land, the latter aspect is also recognizable in the frequency of occurrence of hotspots. Hotspots with identified logistics land are evidently located closer to inland terminals compared to the other grid cells with identified logistics land (see Figure 10). Additionally, they also contain a higher number of existing commercial/industrial land (see Figure 11). Accordingly, the trend of logistical clusterization in a polycentric area, e.g. described in van den Heuvel et al. (2013), can also be identified in the study area at hand.

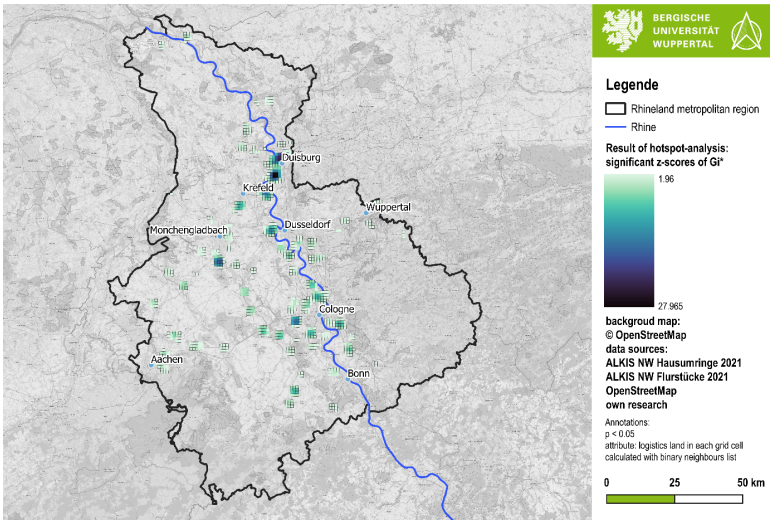


Figure 9: Identified hotspots of logistics land according to G_i^*-statistics by Getis, Ord (1992)

Figure 10: Boxplot for the driving distance to the next inland terminal for grid cells investigated

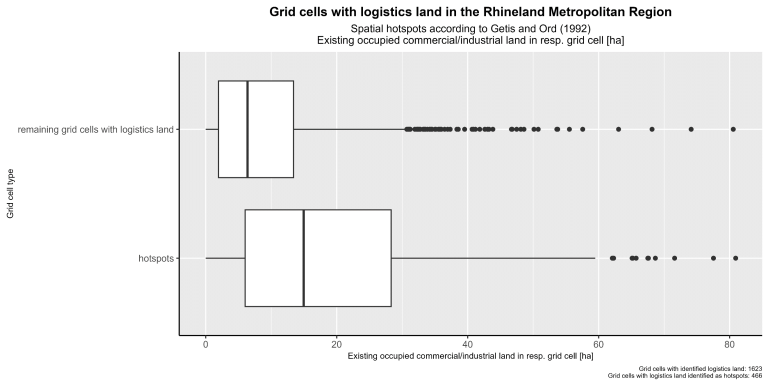


Figure 11: Boxplot for the existing commercial/industrial land in each grid cell

6 Conclusions

This paper presented a method for identifying logistics land based on publicly available data. In addition, area-wide accessibility analyses are carried out. To calculate geostatistical measures (such as Moran's I), all variables considered here were aggregated to grid cells. Accordingly, those presented here represent an image of the existing spatial structure of logistics in the Rhineland metropolitan region. A possible next step would be studies that focus on individual logistics facilities. Further, the generated dataset can also be used to identify underutilized logistics estates and thus potential for densification.

Using hotspot analysis by Getis, Ord (1992), spatial hotspots of logistics land were identified. These hotspots are mostly located in Metropolises/Regiopolises and their suburbs, along highways in areas with flat relief, and in the vicinity to large inland terminals/inland harbors. The results show that – like in other polycentric areas – logistical clusterization can also be observed in the study area at hand.

Further research is needed regarding the following aspects: First, the dataset can be used as a starting point for the observation of the development of logistics facilities over a period of time, as it is part of many other studies like Dablanc et al. (2014). Based on such observations also an evaluation of strategic municipal/regional planning regarding the presence of logistics land would be possible. Additionally, the dataset itself should be validated due to the described regionally inconsistent assignment of the building functions. Further examinations that allow also a comparability to other study areas can be done with an additional differentiation of the logistics facilities according to the typology of logistics facilities like in Heitz et al. (2019).

Bibliography

BA – Federal Employment Agency (2021) Special evaluation of employee data. Employees subject to social insurance at place of work. Unpublished data

Bezirksregierung Köln (2021) Bodenrichtwerte NRW 2021. Web page. Available online at https://www.opengeodata.nrw.de/produkte/infrastruktur_bauen_wohnen/boris/-BRW/BRW_2020_EPSG25832_Shape.zip

BKG – Federal Agency for Cartography and Geodesy (BKG) (2023) Administrative areas as of 01.01. for the territory of the Federal Republic of Germany. Web page. Available online at https://gdz.bkg.bund.de/index.php/default/verwaltungsgebiete-1-250-000-stand-01-01-vg250-01-01.html, updated on 1/1/2023

Busch R (2013) *Logistikimmobilienstandorte in Deutschland – Raumstrukturen und räumliche Entwicklungstendenzen. Eine quantitative Untersuchung mit Hilfe der Baufertigstellungs- und Beschäftigtenstatistik*. MV-Wissenschaft, Münster

Cliff AD, Ord K (1970) Spatial autocorrelation: A review of existing and new measures with applications. *Economic Geography* 46: 269. CrossRef

Dablanc L, Ogilvie S, Goodchild A (2014) Logistics sprawl. *Transportation Research Record: Journal of the Transportation Research Board* 2410: 105–112. CrossRef

Dablanc L, Rakotonarivo D (2010) The impacts of logistics sprawl: How does the location of parcel transport terminals affect the energy efficiency of goods' movements in Paris and what can we do about it? *Procedia - Social and Behavioral Sciences* 2: 6087–6096. CrossRef

Geobasis NRW (2020) WFS NW ALKIS Grundrissdaten vereinfachtes Schema. Available online at https://www.wfs.nrw.de/geobasis/wfs_nw_alkis_vereinfacht?SERVICE=-WFS&VERSION=2.0.0&REQUEST=GetCapabilities

Geobasis NRW (2021) Hausumringe NW. Available online at https://open.nrw/dataset/-3f08a580-48ec-43c1-936d-d62f89c21cc9, checked on 1/25/2022

Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189–206. CrossRef

Heitz A, Beziat A (2016) The parcel industry in the spatial organization of logistics activities in the Paris region: Inherited spatial patterns and innovations in urban logistics systems. *Transportation Research Procedia* 12: 812–824. CrossRef

Heitz A, Dablanc L, Tavasszy LA (2017) Logistics sprawl in monocentric and polycentric metropolitan areas: The cases of Paris, France, and the Randstad, the Netherlands. *REGION* 4: 93. CrossRef

Heitz A, Launay P, Beziat A (2019) Heterogeneity of logistics facilities: An issue for a better understanding and planning of the location of logistics facilities. *European Transport Research Review* 11. CrossRef

Holguín-Veras J, Lawson C, Wang C, Jaller M, González-Calderón C, Campbell S, Kalahashti L, Wojtowicz J, Ramirez D (2016) *Using Commodity Flow Survey Microdata and Other Establishment Data to Estimate the Generation of Freight, Freight Trips, and Service Trips: Guidebook*. Transportation Research Board. CrossRef

Holguín-Veras J, Wang C, Ng J, Ramírez-Ríos D, Wojtowicz J, Calderón O, Caron B, Rivera-González C, Pérez S, Schmid J, Kim W, Ismael A, Amaral JC, Lawson C, Haake D (2022) *Planning Freight-Efficient Land Uses: Methodology, Strategies, and Tools*. Transportation Research Board. CrossRef

Holl A, Mariotti I (2017) The geography of logistics firm location: The role of accessibility. *Networks and Spatial Economics* 18: 337–361. CrossRef

IT.NRW – Landesbetrieb Information und Technik Nordrhein-Westfalen (2021) Sonderauswertung des Unternehmensregistersystems. Anzahl der Niederlassungen in den Gemeinden von NRW in 2006 - 2019 in ausgewählten WZ-Gruppen. Unpublished data

Jaller M, Zhang X, Qian X (2022) Distribution facilities in California: A dynamic landscape and equity considerations. *Journal of Transport and Land Use* 15. CrossRef

Kirilenko AP (2022) Geographic information system (GIS). In: Egger R (ed), *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*. Springer International Publishing, Cham, 513–526. CrossRef

Klauenberg J, Krause Cauduro F (2020) Logistics locations patterns—distance-based methods for relative industrial concentration measurement applied to the region of Berlin–Brandenburg. In: Elbert R, Friedrich C, Boltze M, Pfohl HC (eds), *Urban Freight Transportation Systems*. Elsevier, 3–18. CrossRef

Kretzschmar D, Gutting R, Schiller G, Weitkamp A (2021) Warenlagergebäude in Deutschland: Eine neue Methodik zur regionalen Quantifizierung der Flächeninanspruchnahme. *Raumforschung und Raumordnung* 79: 136–153. CrossRef

Leerkamp B, Thiemermann A, Groß F, Holthaus T, Jaeger A, Janßen T, Stock S, Oralek F, Janßen L, Siefer T, Sewcyk B, Busch R (2022) Güterverkehrsstudie für das Gebiet der Metropolregion Rheinland. Endbericht. Final report, edited by Nahverkehr Rheinland (NVR). Available online at https://wir.gorheinland.com/vernetzte-mobilitaet/regionale-konzepte/gueterverkehrsstudie/, updated on 9/1/2022, checked on 12/23/2024.

Nefs M, Zonneveld W, Daamen T, van Oort F (2024) Landscapes of trade: Towards sustainable spatial planning for the logistics complex in the Netherlands. Thesis, Delft University of Technology

OSM – OpenStreetMap contributors (2021) Gebäudedatensatz Metropolregion Rheinland. Available online at https://openstreetmap.org, updated on 1/18/2021

O'Sullivan D, Unwin DJ (2010) *Geographic Information Analysis*. Wiley. CrossRef

Sakai T, Beziat A, Heitz A (2020) Location factors for logistics facilities: Location choice modeling considering activity categories. *Journal of Transport Geography* 85: 102710. CrossRef

Sakai T, Kawamura K, Hyodo T (2015) Locational dynamics of logistics facilities: Evidence from Tokyo. *Journal of Transport Geography* 46: 10–19. CrossRef

Trent NM, Joubert JW (2022) Logistics sprawl and the change in freight transport activity: A comparison of three measurement methodologies. *Journal of Transport Geography* 101: 103350. CrossRef

van den Heuvel FP, de Langen PW, van Donselaar KH, Fransoo JC (2013) Spatial concentration and location dynamics in logistics: The case of a Dutch province. *Journal of Transport Geography* 28: 39–48. CrossRef

Funded by

ersa

WU

FWF