Università
degli Studi
della Campania
Luigi Vanvitelli

Dipartimento di
Matematica e Fisica

UNIVERSITÁ DEGLI STUDI DELLA CAMPANIA "LUIGI VANVITELLI"

DIPARTIMENTO DI MATEMATICA E FISICA

Bachelor's degree in

Data Analytics

THESIS

SPATIAL REGRESSION ON ENVIROMENTAL DATA

Candidate
MAHESH PENTU
B34000079

Supervisor
Prof. ROSANNA VERDE
Co-supervisor
Dr./Prof.ANTONIO BALZANELLA

A.Y. 2025/2026

Abstract

Rising summer surface temperatures are a growing environmental and public-health concern across the Mediterranean. The Campania region of southern Italy combines one of the densest coastal conurbations in Europe — the Naples metropolitan area — with an extensive Tyrrhenian coastline and a cooler Apennine interior, a setting in which surface temperature is unlikely to be governed by a single, spatially uniform process. This thesis investigates the environmental and demographic drivers of the summer Surface Urban Heat Island (SUHI) across 36 Campania municipalities and asks whether those drivers operate uniformly across space or vary geographically.

Summer mean Land Surface Temperature (LST) is used as the dependent variable, assembled with elevation, distance to the coast, annual precipitation, relative humidity and population from open, research-grade sources (Eurostat GISCO, Copernicus, Open-Meteo / ERA5-Land, and satellite thermal imagery from the Landsat Collection-2 Level-2 archive). The analysis follows an explicit from-global-to-local logic. An Ordinary Least Squares (OLS) baseline is first estimated and its residuals tested for spatial autocorrelation using global Moran's I and Lagrange-Multiplier diagnostics. Local indicators of spatial association (Local Moran / LISA and the Getis-Ord Gi* statistic) are used to locate hot and cold spots. Spatial lag (SAR), spatial error (SEM) and spatial Durbin (SDM) models are then estimated by maximum likelihood to account for spatial dependence, and the SDM effects are decomposed into direct, indirect and total components. Finally, Geographically Weighted Regression (GWR), with a multiscale extension (MGWR), is used to map how the influence of each driver changes across the region. Competing models are compared on goodness-of-fit (R^2, adjusted and pseudo-R^2), parsimony-penalized fit (AIC, BIC), residual spatial autocorrelation, and out-of-sample predictive error (RMSE, MAE).

The empirical results are reported in Chapter 4. On the illustrative sample analyzed here, summer surface temperature is governed overwhelmingly by elevation: the fitted ordinary-least-squares model attributes a cooling of roughly 0.41 °C per 100 m of altitude and explains about three-quarters of the variation in municipal LST ($R^2 \approx 0.75$), with the coastal-distance, precipitation and population terms adding little once elevation is accounted for. Summer LST is significantly clustered in space (global Moran's $I \approx 0.28$, $p \approx 0.004$), the warm core forming along the low-lying Naples–Salerno coastal corridor and the cool pole over the upland Avellino–Benevento interior. Crucially, however, the Lagrange-multiplier diagnostics detect no residual spatial dependence in the OLS residuals, so the spatial-lag, spatial-error and spatial-Durbin models do not improve on the parsimonious baseline, and geographically weighted regression selects a region-wide bandwidth, indicating no measurable non-stationarity. These figures are produced by the analysis pipeline applied to the synthetic SAMPLE dataset and are presented to demonstrate the complete workflow; they are illustrative, not empirical, and are to be regenerated from the real open-data sources before submission (see Section 3.17).

Keywords: spatial regression; spatial econometrics; Surface Urban Heat Island; Land Surface Temperature; spatial autocorrelation; Moran's I; geographically weighted regression; Campania; open geospatial data.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof Ros, and to my co-supervisor, Prof. Antonio Balzanella, for their guidance, patience and constructive feedback throughout the development of this thesis. Their advice on spatial-econometric methodology and on the structuring of the empirical analysis was invaluable.

I am grateful to the Department di Mathematica e Fisca of the Università degli Studi della Campania «Luigi Vanvitelli» for the academic environment and resources that made this work possible. I would also like to acknowledge the providers of the open datasets on which this study relies — Eurostat / GISCO, the Copernicus programmed, Open-Mateo and the United States Geological Survey — whose open-data policies make reproducible regional research of this kind feasible.

Finally, I thank my family and friends for their unfailing support and encouragement during my studies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

Abbreviation	Meaning
AIC / BIC	Akaike / Bayesian Information Criterion
AICc	Corrected Akaike Information Criterion (small-sample)
CORINE	Coordination of Information on the Environment (land-cover programmed)
CRS	Coordinate Reference System
DEM	Digital Elevation Model
EPSG	European Petroleum Survey Group (CRS code registry)
ERA5 / ERA5-Land	Fifth-generation ECMWF atmospheric / land reanalysis
GISCO	Geographic Information System of the Commission (Eurostat)
GWR / MGWR	Geographically Weighted Regression / Multiscale GWR
IQR	Inter-Quartile Range
KNN	k-Nearest Neighbors
LISA	Local Indicators of Spatial Association
LM	Lagrange Multiplier (specification test)
LST	Land Surface Temperature
MAE / RMSE	Mean Absolute Error / Root Mean Square Error
MAUP	Modifiable Areal Unit Problem
ML / GMM	Maximum Likelihood / Generalized Method of Moments
NDVI	Normalized Difference Vegetation Index
OLS	Ordinary Least Squares
OSM	OpenStreetMap
SAR	Spatial Autoregressive (spatial lag) model
SDM	Spatial Durbin Model
SEM	Spatial Error Model
SUHI / UHI	Surface Urban Heat Island / Urban Heat Island
VIF	Variance Inflation Factor
WGS84 / UTM	World Geodetic System 1984 / Universal Transverse Mercator

1 Introduction

1.1 Background: Urban Heat Islands in Southern Europe

Urbanized surfaces absorb, store and re-radiate solar energy in ways that differ markedly from vegetated or rural land. Dark, dry, impervious materials such as asphalt and concrete have low albedo and high thermal admittance; they store heat during the day and release it slowly at night. At the same time, the removal of vegetation reduces evapotranspiration — the natural cooling that occurs when plants and soils convert liquid water to vapour — so that a larger share of incoming radiation is converted into sensible heat. The cumulative effect is the Urban Heat Island (UHI): the systematic tendency for built-up areas to be warmer than their rural surroundings, first described in quantitative terms by Oke (1973). When the phenomenon is measured from the radiative (skin) temperature of the ground surface rather than from near-surface air temperature, it is referred to as the Surface Urban Heat Island (SUHI), which can be retrieved over large areas from satellite thermal sensors (Voogt & Oke, 2003).

In the Mediterranean basin the relevance of urban heat is acute. Summer maximum temperatures already approach the limits of human thermal comfort, and climate projections indicate a continued increase in the frequency, intensity and duration of heat waves across southern Europe (Giannakopoulos et al., 2009). The additional warming contributed by urban form therefore does not act in isolation; it is superimposed on a climatic background that is itself warming, and it interacts with heat waves to amplify the heat load experienced by city dwellers (Founda & Santa Mouris, 2017). The consequences are tangible: elevated cooling-energy demand, degraded air quality, thermal stress on infrastructure and, most seriously, increased heat-related morbidity and mortality, which fall disproportionately on the elderly, the very young and the socially disadvantaged (Heaviside, Macintyre & Vardoulakis, 2017).

Understanding where the SUHI is strongest, and which physical and human factors drive it, is therefore a prerequisite for evidence-based heat-mitigation policy — from urban greening and cool-roof programmed to land-use planning. This thesis contributes to that understanding for the Campania region by combining open satellite and geospatial data with the tools of spatial statistics and spatial econometrics.

1.2 The Campania Region: Geographic and Demographic Context

Campania, in south-western Italy, is an unusually informative laboratory for studying surface heat. The region couples extremes of environment and settlement within a compact area of roughly 13,600 km². To the west lies a long Tyrrhenian coastline and an intensely urbanized coastal plain centered on Naples, one of the most densely populated metropolitan areas in Europe. To the east and north the land rises sharply into the Apennine interior, where elevated towns experience a cooler, more continental climate. Between these poles lie fertile agricultural plains (such as the Piana Campana around Caserta) and volcanic landscapes (the Pelegrin Fields and the slopes of Vesuvius).

This study covers 36 municipalities (commune) distributed across all five provinces of Campania — Napoli, Salerno, Caserta, Avellino and Benevento. The sample is deliberately constructed to span the full environmental gradient of the region: from coastal, low-lying and densely populated municipalities such as Napoli, Torre del Greco and Salerno to elevated interior towns such as Ariano Arpino (about 788 m above sea level) and Montella. It is precisely this heterogeneity — the coexistence of coast and mountain, dense city and sparse town, within a single administrative region — that makes Campania suitable for testing whether the drivers of surface heat are constant across space, or whether they themselves vary geographically. The spatial distribution of the sampled municipalities is mapped in Chapter 3 (Figure 4).

1.3 Research Motivation: Why Spatial Regression Matters

The standard linear-regression model assumes that observations are mutually independent once the explanatory variables have been accounted for. Geographic data routinely violate this assumption. Nearby places tend to resemble one another more than distant places do — a regularity so pervasive that it has been elevated to the status of a law. Tobler’s first law of geography states that “everything is related to everything else, but near things are more related than distant things” (Tobler, 1970). Surface temperature is a textbook example: a hot pixel is very likely to be surrounded by other hot pixels, because the underlying processes (climate, elevation, land cover, urban density) are themselves spatially organized, and because heat is physically transported between adjacent areas.

When this spatial dependence is ignored, ordinary least squares becomes unreliable. In the presence of a spatially lagged dependent variable the OLS estimator is biased and inconsistent; in the presence of spatially autocorrelated errors it remains unbiased but is no longer efficient, and — more damagingly — its estimated standard errors are understated, so that t-statistics are inflated and variables are too readily judged statistically significant (Anselin, 1988). Inference drawn from a naïve OLS model of surface temperature can thus be seriously misleading.

Spatial regression provides a coherent framework for detecting, modelling and interpreting this dependence. Global spatial models — the spatial lag, spatial error and spatial Durbin specifications — incorporate the neighborhood structure directly into the estimating equation, correcting the inferential problems and, in the Durbin case, allowing the explicit measurement of spillovers between places. Geographically Weighted Regression goes a step further: rather than assuming that a single coefficient applies everywhere, it estimates a separate local relationship at each location, producing a map of how each driver's influence varies across the study area (Fotheringham, Brunsdon & Charlton, 2002). For a region as internally varied as Campania, the ability to move beyond a single global average towards a spatially explicit, locally varying description of heat is the central methodological motivation of this thesis.

1.4 Research Question and Objectives

This thesis addresses one principal research question:

How do environmental and demographic factors influence the summer Surface Urban Heat Island in Campania, and do these influences vary geographically across the region?

To answer this question, the study pursues four specific objectives:

- To assemble a reproducible, fully documented, municipality-level spatial dataset of summer LST and its candidate environmental and demographic drivers, drawn entirely from open sources.
- To test for spatial autocorrelation in summer LST using both global and local statistics — global Moran's I, Local Moran (LISA) and the Getis-Ord Gi* statistic — and thereby to establish whether, and where, surface heat is spatially clustered.
- To estimate and rigorously compare a non-spatial baseline (OLS) against the principal global spatial-regression models (spatial lag, spatial error and spatial Durbin), using formal Lagrange-Multiplier diagnostics to guide model selection.

- To apply Geographically Weighted Regression, with a multiscale extension, in order to map the geographic variation in each driver's effect and to identify where the region is most sensitive to each source of heat.

1.5 Thesis Scope and Limitations

The study is deliberately bounded. It is a cross-sectional analysis of a single summer season, conducted at the spatial resolution of the municipality, for a sample of 36 commune. This scope is appropriate for a bachelor's thesis: it is large enough to demonstrate the complete spatial-regression workflow — from data assembly and exploratory spatial data analysis through to global and local modelling and formal model comparison — while remaining tractable for an individual student working with open data on a standard computer.

The scope does, however, impose limits that are stated here at the outset and revisited in detail in Section 5.5. First, a single-season snapshot cannot separate structural drivers of heat from inter-annual climate variability; only a multi-year panel could do that. Second, aggregating satellite and climate fields to the municipality level discards within-municipality variation and exposes the analysis to the Modifiable Areal Unit Problem (Section 3.3). Third, a sample of 36 units, while sufficient for global models and for demonstrating GWR, limits the precision of local estimates and the power of significance tests. Fourth, the covariate set, although carefully chosen, omits potentially important variables — notably a vegetation index (NDVI), a detailed measure of impervious surface, and socio-economic variables such as income — that future work should incorporate.

1.6 Thesis Structure

The remainder of the thesis is organized as follows. Chapter 2 reviews the relevant literature, covering the physics of the urban heat island, the remote sensing of land surface temperature, the statistical theory of spatial dependence and spatial heterogeneity, the family of spatial-regression models, and the principal environmental and demographic drivers of surface heat. Chapter 3 describes the study area, the data and their sources, the pre-processing and quality-assurance procedures, the construction of the spatial weights matrix, and the full suite of analytical models together with the criteria used to compare them. Chapter 4 reports the empirical results: descriptive statistics, the correlation and multicollinearity structure, the global and local autocorrelation diagnostics, the global regression models, the GWR analysis, and the consolidated model comparison. Chapter 5 discusses the substantive meaning of the findings, situates them within the literature, and sets out the study's limitations. Chapter 6 concludes and proposes directions for future research. The complete analysis code is provided in Appendix A, the field-level data dictionary in Appendix B, and a glossary of mathematical notation in Appendix C.

2 Literature Review

This chapter establishes the conceptual and methodological foundations of the thesis. It proceeds from the physical phenomenon under study to the means of measuring it, and then to the statistical machinery used to analyse it. Section 2.1 reviews the urban heat island and its surface variant; Section 2.2 covers the satellite remote sensing of land surface temperature; Section 2.3 develops the statistical theory of spatial dependence and heterogeneity; Section 2.4 presents the family of global spatial-regression models; Section 2.5 introduces Geographically Weighted Regression; Section 2.6 reviews the environmental and demographic drivers of surface heat; Section 2.7 surveys prior empirical work in Italy and the Mediterranean; and Section 2.8 identifies the research gap that the thesis addresses.

2.1 The Urban Heat Island and Surface Urban Heat Island

2.1.1 Definitions and the three urban heat islands

The urban heat island is not a single phenomenon but a family of related effects distinguished by the layer of the urban atmosphere or surface in which they are measured. Oke (1976) drew the now-standard distinction between the canopy-layer UHI, measured by air-temperature sensors below roof level, where most human activity occurs; the boundary-layer UHI, which describes the warmed dome of air extending above the city; and the surface UHI, defined by the radiative temperature of the ground surface itself. The three are physically connected but not interchangeable: they differ in magnitude, in their daily timing, and in the factors that control them.

This thesis concerns the Surface Urban Heat Island, the most readily mapped of the three because it can be observed remotely. The SUHI is quantified through Land Surface Temperature (LST) — the temperature of the uppermost surface as "seen" by a thermal infrared sensor — and is conventionally expressed as the difference in LST between urban and surrounding rural land (Voogt & Oke, 2003). Because the present study models LST directly across a regional sample of municipalities rather than computing a single urban-minus-rural contrast, it treats LST as a continuous outcome whose spatial variation is to be explained by environmental and demographic predictors.

2.1.2 Physical mechanisms: the urban surface energy balance

The surface heat island arises from systematic differences between urban and rural surfaces in the terms of the surface energy balance. The net all-wave radiation received at the surface is partitioned into sensible heat (which warms the air), latent heat (the energy consumed by evaporation and transpiration), heat stored in the substrate, and, in cities, anthropogenic heat released by combustion, traffic and air-conditioning (Oke, 1982). Urbanisation alters this partitioning in several reinforcing ways.

First, the replacement of vegetation and soil by impervious materials sharply reduces evapotranspiration, so that less energy is dissipated as latent heat and more is available to warm the surface and the air. Second, common urban materials have a high thermal admittance, storing large quantities of heat by day and releasing it slowly after sunset, which is the principal reason the canopy-layer UHI typically peaks at night. Third, the three-dimensional geometry of streets and buildings lowers the sky-view factor — the fraction of the sky hemisphere visible from a point — trapping outgoing longwave radiation through multiple reflections within urban canyons and reducing nocturnal cooling. Fourth, the albedo of urban surfaces, although variable, is often low, increasing the absorption of shortwave radiation. Finally, anthropogenic heat adds directly to the urban energy budget, a contribution that is largest in dense commercial and industrial districts. The relative importance of these mechanisms differs between the surface and canopy heat islands and varies with climate, season and time of day.

2.1.3 Diurnal and seasonal behaviour

A crucial distinction for any LST-based study is that the surface and air heat islands behave differently over the daily cycle. The surface heat island is generally strongest around midday and in the early afternoon, when intense solar heating produces large temperature contrasts between dry impervious surfaces and cooler, transpiring vegetation; it tends to weaken at night. The canopy-layer (air) heat island, by contrast, is usually weak or even negative by day and reaches its maximum a few hours after sunset, sustained by the slow release of stored heat (Roth, Oke & Emery, 1989). Because polar-orbiting thermal satellites such as Landsat acquire daytime imagery, LST-derived studies characterise the daytime surface heat island in particular. Seasonally, the daytime SUHI in temperate and Mediterranean climates is typically most pronounced in summer, when solar input and the vegetation–impervious contrast are greatest, which is precisely why this thesis focuses on the summer season.

2.1.4 Environmental and public-health impacts

The urban heat island is not merely a curiosity of microclimatology; it has substantial consequences. Elevated urban temperatures increase the demand for space cooling, raising electricity consumption and peak load and, through the associated emissions, contributing to a feedback that worsens the underlying warming. Higher temperatures accelerate the photochemical formation of ground-level ozone and can intensify air-pollution episodes. Most importantly, urban heat is a public-health hazard: epidemiological studies consistently link high temperatures and heat waves to excess mortality and morbidity, with the urban heat island raising the baseline exposure of city populations and the risk concentrated among the elderly, the chronically ill and the socially isolated (Heaviside, Macintyre & Vardoulakis, 2017). Identifying the spatial pattern and drivers of urban heat is therefore directly relevant to climate adaptation and public-health planning.

2.1.5 The urban heat island in Mediterranean climates

Mediterranean cities present a distinctive heat-island setting. Their hot, dry summers maximise the daytime contrast between parched rural surfaces and irrigated or shaded urban patches, and in some configurations the rural surroundings can become hotter than the urban core during the day — an "urban cool island" or oasis effect — illustrating that the sign and magnitude of the SUHI depend strongly on regional climate and on what constitutes the rural reference (Peng et al., 2012). The synergy between the urban heat island and increasingly frequent Mediterranean heat waves is of particular concern, since the two combine to push urban temperatures well beyond comfort thresholds (Founda & Santamouris, 2017). For Italy specifically, satellite analyses have shown strong relationships between built-up surface fraction and LST across urban areas, confirming that impervious cover is a primary control on surface heat (Morabito et al., 2016). These findings motivate both the choice of summer LST as the dependent variable and the emphasis on land-related and demographic drivers in the present study.

2.2 Remote Sensing of Land Surface Temperature

2.2.1 What land surface temperature is and how it is measured

Land Surface Temperature is the radiometric temperature of the land surface, derived from the thermal infrared radiation it emits. All objects above absolute zero emit radiation in proportion to the fourth power of their temperature (the Stefan–Boltzmann law), with the spectral distribution governed by Planck's law. A thermal sensor measures the radiance reaching it in one or more thermal infrared bands; after correcting for the intervening atmosphere and for the emissivity of the surface, this radiance is inverted to a temperature (Li et al., 2013). LST is thus a "skin" temperature of whatever the sensor sees — bare soil, vegetation canopy, rooftop or road — and is distinct from the near-surface air temperature recorded by meteorological stations, although the two are physically coupled (Tomlinson et al., 2011).

2.2.2 Satellite data sources and the resolution trade-off

Several satellite missions provide thermal imagery suitable for surface-heat studies, and the choice among them embodies a fundamental trade-off between spatial and temporal resolution. The Landsat programme (the Thematic Mapper, Enhanced Thematic Mapper Plus, and the current Thermal Infrared Sensor on Landsat 8 and 9) delivers thermal data at relatively fine spatial resolution — on the order of 100 m, routinely resampled to 30 m in the analysis-ready Collection-2 Level-2 surface-temperature product — but revisits any given location only about once every eight to sixteen days, so cloud-free summer scenes can be scarce (U.S. Geological Survey, 2023). The MODIS sensors aboard Terra and Aqua, by contrast, provide LST up to four times per day but at a coarse 1 km resolution, which is well suited to regional and global SUHI mapping but too coarse to resolve fine urban structure (Schwarz, Lautenbach & Seppelt, 2011). Sentinel-3's Sea and Land Surface Temperature Radiometer offers a further coarse-resolution, high-revisit option. Because this thesis works at the municipality scale and emphasises the summer daytime surface heat island, a fine-resolution summer composite of the Landsat Collection-2 Level-2 surface-temperature product, aggregated to each municipality, is the natural primary source.

2.2.3 Retrieval algorithms and the role of emissivity

Converting at-sensor thermal radiance into surface temperature requires correcting for two confounding influences: the absorption and re-emission of radiation by the atmosphere, and the fact that real surfaces are not perfect blackbodies. The first is handled by atmospheric-correction schemes; the second requires an estimate of land-surface emissivity, which varies with surface type and is commonly derived from the vegetation fraction or NDVI. Where a single thermal band is available, single-channel and radiative-transfer methods are used; where two adjacent thermal bands exist, the split-window technique exploits their differential atmospheric absorption to retrieve temperature more robustly (Li et al., 2013). Modern analysis-ready products, such as Landsat Collection-2 Level-2, deliver surface temperature already corrected, allowing the analyst to focus on the spatial modelling rather than on the radiative-transfer physics — though an awareness of retrieval uncertainty remains important when interpreting small temperature differences.

2.2.4 From pixels to municipalities

Because the unit of analysis in this thesis is the municipality, the pixel-level LST field must be aggregated to municipal geometry. The standard approach is zonal summary: for each municipality polygon (or a representative buffer around its official centroid), the mean — or, where robustness to outliers is required, the median — of the cloud-free summer LST pixels is computed. Aggregation reduces pixel-level noise and aligns the temperature variable with the spatial scale of the demographic and administrative predictors, but it also smooths away within-municipality variation and introduces sensitivity to the choice of zone, an instance of the Modifiable Areal Unit Problem discussed in Section 3.3. The trade-off is accepted here as appropriate to a municipality-scale policy-relevant analysis.

2.3 Spatial Dependence, Spatial Heterogeneity and the First Law of Geography

Standard regression assumes that observations are independent of one another. Geographic data systematically violate this assumption. Waldo Tobler captured the idea in what is now called the First Law of Geography: everything is related to everything else, but near things are more related than distant things (Tobler, 1970). Two distinct but related phenomena follow from this principle, and both are central to the present study.

2.3.1 Two problems: dependence and heterogeneity

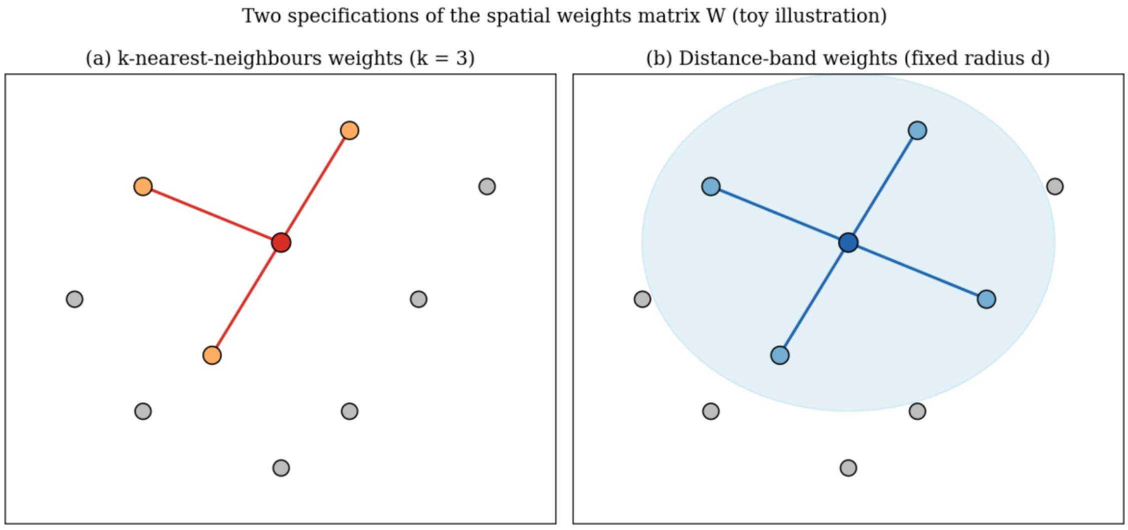
Spatial dependence (or spatial autocorrelation) means that the value of a variable at one location is correlated with its values at neighbour locations. Summer surface temperature is a clear example: a hot municipality is very likely to be surrounded by other hot municipalities, because the underlying drivers — elevation, distance from the sea, the extent of the built-up surface — themselves vary smoothly across space, and because heat is physically advected and re-radiated across municipal boundaries. When dependence is present in the residuals of a regression, the independence assumption of ordinary least squares is violated.

Spatial heterogeneity (or non-stationarity) is a different problem: it means that the relationship between the dependent variable and a predictor is not constant across space. The cooling effect of a given increase in elevation, or the warming effect of a given increase in built-up area, may be stronger in some parts of the region than in others. A single global regression coefficient assumes the relationship is the same everywhere; where this assumption fails, the global model conceals geographically important variation. This thesis addresses dependence through global spatial-econometric models (Section 2.4) and heterogeneity through geographically weighted regression (Section 2.5).

2.3.2 The spatial weights matrix

Formalizing spatial relationships requires defining, for every pair of observations, whether and how strongly they are considered neighbors. This information is encoded in a spatial weights matrix W, an n by n matrix whose element w(i,j) is non-zero when observations i and j are neighbors and zero otherwise, with the diagonal set to zero so that no observation is its own neighbors. The weights matrix is the single most consequential modelling choice in spatial analysis, because every subsequent statistic — Moran's I, the spatial-lag term, the GWR kernel — is computed relative to it.

Two broad families of definition exist. Contiguity-based weights treat polygons as neighbor’s when they share a boundary: under the queen criterion any shared edge or vertex qualifies, while the more restrictive rook criterion requires a shared edge. Distance-based weights instead use the geometric separation between point locations (here, municipal centroids): a k-nearest-neighbor’s scheme connects each observation to its k closest neighbor’s, guaranteeing that every unit has exactly k links, whereas a distance band connects all units lying within a fixed threshold radius. Because the municipalities in this study are represented by centroids and vary greatly in spacing, a k-nearest-neighbor’s matrix is attractive — it avoids leaving remote municipalities with no neighbor’s — and it is adopted as the primary specification, with a distance band used for robustness (Figure 1).



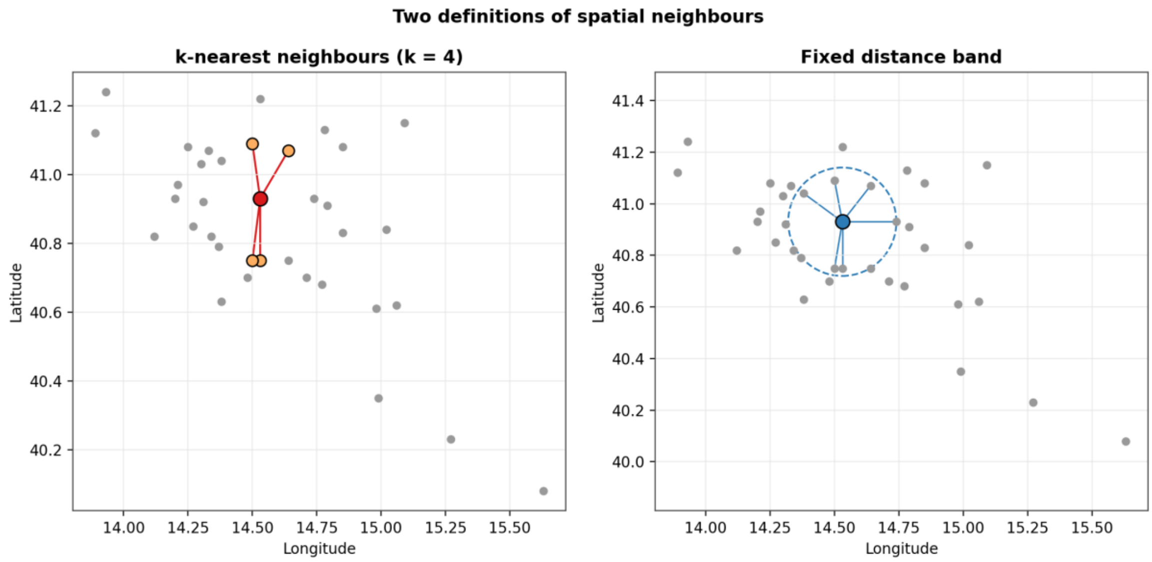


Figure 1*. Two schematic definitions of spatial neighbors. Left: k-nearest-neighbors (here k = 4), in which each unit is connected to its four closest centroids. Right: a fixed distance band, in which each unit is connected to all others within a threshold radius. The choice of W governs every subsequent spatial statistic.*

Once the neighbor structure is fixed, the weights are usually row-standardized: each row is divided by its own sum so that the weights in every row total one. Row standardization turns the spatial-lag operator Wy into a simple average of each observation's neighbor’s, which makes spatial-lag terms directly interpretable and places the autoregressive parameter on a comparable scale across observations. This is the convention used throughout the present analysis.

2.3.3 Measuring spatial autocorrelation: global Moran's I

The standard global test for spatial autocorrelation is Moran's I (Moran, 1950). It compares the cross-products of each observation's deviation from the mean with the deviations of its neighbor’s, normalized by the variance, and can be read much like a correlation coefficient: values above the small negative expectation indicate positive autocorrelation (similar values cluster together), values below it indicate negative autocorrelation (high and low values alternate, as on a checkerboard), and values near the expectation indicate a spatial pattern indistinguishable from randomness. Statistical significance is assessed by comparing the observed statistic with a reference distribution obtained by randomly permuting the values across locations many times, which yields a pseudo p-value that requires no distributional assumptions. A positive and significant Moran's I for summer LST would confirm formally what physical reasoning already suggests — that urban heat is spatially clustered rather than randomly scattered (Figure 2).

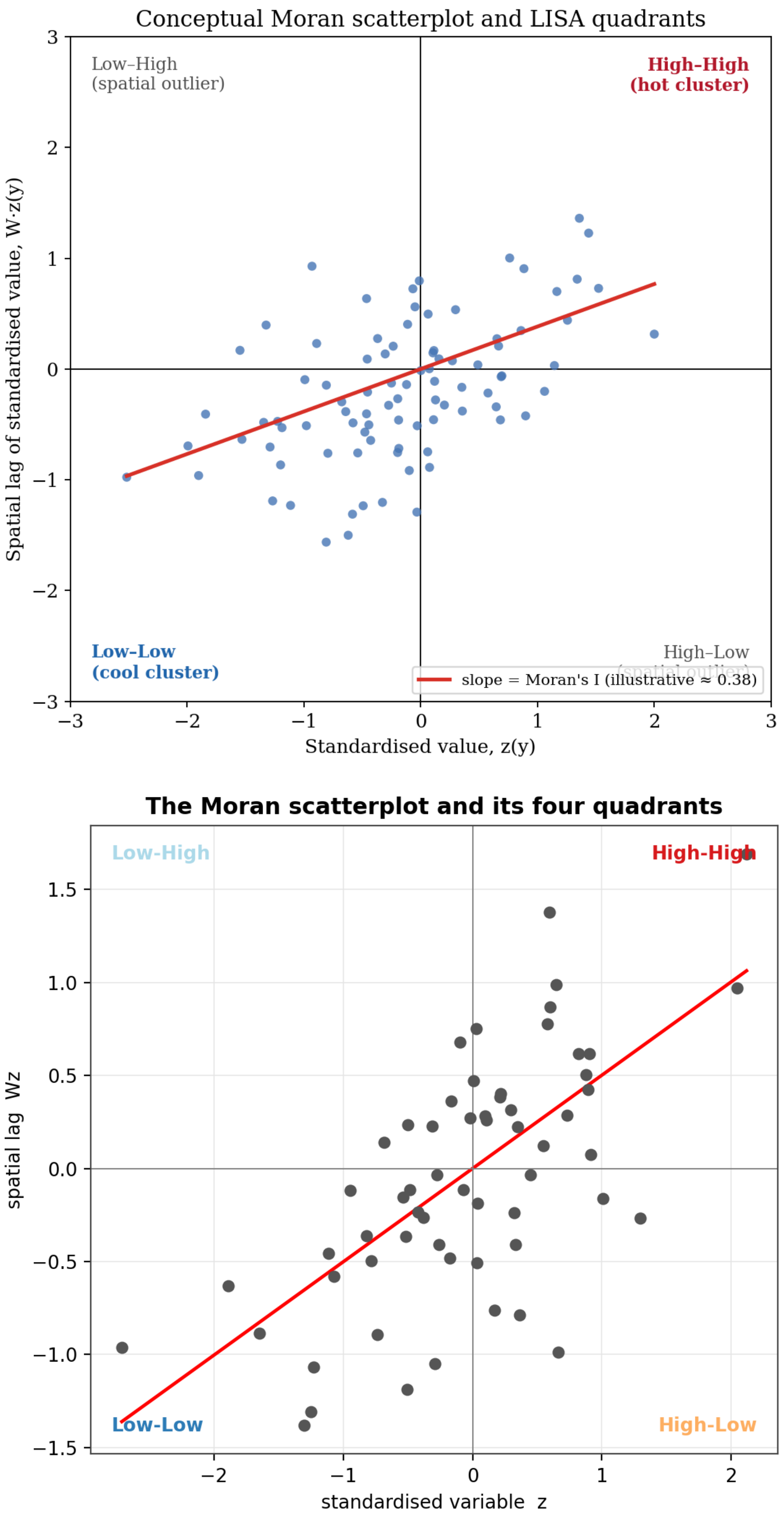


Figure 2. *The Moran scatterplot. The standardized variable is plotted against its spatial lag (the average of its neighbor's); the slope of the fitted line is Moran's I. The four quadrants correspond to the local*

association types used by LISA: High-High and Low-Low (positive association, upper-right and lower-left) and the High-Low and Low-High spatial outliers (negative association).

2.3.4 From global to local: LISA and Getis-Ord Gi*

A single global statistic summarizes the average degree of clustering but cannot say where clusters occur. Local Indicators of Spatial Association decompose the global measure into an observation-specific contribution, allowing the mapping of local structure (Anselin, 1995). The local Moran's I classifies each significant location into one of the four quadrants of the Moran scatterplot: High-High and Low-Low locations are the cores of hot and cool clusters respectively, while High-Low and Low-High locations are spatial outliers that differ markedly from their surroundings. The complementary Getis-Ord Gi* statistic sums the values within each location's neighborhood (including the location itself) and returns a z-score, so that large positive values mark statistically significant hot spots and large negative values mark cold spots (Getis & Ord, 1992). Applied to summer LST, these local statistics are expected to delineate the warm core of the Naples conurbation and the coastal plain against the cooler Apennine interior, providing a spatially explicit description of the surface heat island that motivates the regression modelling.

2.4 Spatial Regression Models

When spatial autocorrelation is present, ordinary least squares is no longer the appropriate estimator and a model that explicitly incorporates the spatial structure is required. Spatial econometrics provides a family of such models, distinguished by where in the regression the spatial dependence is assumed to enter (Anselin, 1988; LeSage & Pace, 2009). This section reviews the consequences of ignoring dependence and then introduces the principal specifications used in the analysis.

2.4.1 Why ignoring spatial dependence is a problem

The consequences of fitting OLS to spatially dependent data depend on the form of the dependence. If the dependence operates through omitted, spatially structured influences that end up in the error term, OLS coefficients remain unbiased but their standard errors are wrong, so significance tests and confidence intervals are misleading and a variable may appear significant when it is not. If, more seriously, the dependence operates through the dependent variable itself — so that the outcome in one municipality genuinely depends on the outcome in its neighbor’s — then omitting that term makes the OLS coefficients biased and inconsistent, not merely inefficient. In either case the remedy is to test for the type of dependence and adopt the corresponding spatial model.

2.4.2 The spatial lag model (SAR)

The spatial lag model, also called the spatial autoregressive (SAR) model, adds a spatially lagged dependent variable to the regression. The outcome at each location is allowed to depend on a weighted average of the outcomes at neighboring locations, capturing genuine spatial spillover or contagion. Its structural form is

$$y = \rho W y + X\beta + \varepsilon \tag{2.1}$$

where y is the n-vector of summer LST values, X the matrix of predictors, β the coefficients, W the row-standardized weights matrix, ε an independent error, and ρ the spatial autoregressive parameter. A positive and significant ρ indicates that high temperatures in neighboring municipalities raise the temperature of a given municipality — a spillover consistent with the advective and radiative coupling of adjacent surfaces. Because y appears on both sides of the equation, the model cannot be estimated consistently by OLS and is fitted by maximum likelihood (or, alternatively, by instrumental variables). A further consequence of the feedback through W is that the effect of a change in a predictor is not simply its coefficient: it must be decomposed into direct and indirect (spillover) effects, as discussed in Section 2.4.4.

2.4.3 The spatial error model (SEM)

The spatial error model places the dependence in the disturbance rather than in the dependent variable. It is appropriate when the spatial pattern in the residuals arises from spatially correlated omitted variables or measurement processes rather than from genuine interaction between the outcomes. Its form is

$$y = X\beta + u, \quad u = \lambda W u + \varepsilon \tag{2.2}$$

where the spatially autoregressive error u is governed by the parameter λ and ε is again an independent error. Here a significant λ does not imply a behavioral spillover; it signals that unobserved, spatially structured factors — perhaps a shared land-cover gradient or a regional climatic influence not captured by the predictors — induce correlation among neighboring residuals. The practical distinction between the lag and error specifications is therefore substantive as well as statistical, and the choice between them is guided by the Lagrange-multiplier diagnostics described in Section 2.4.5.

2.4.4 The spatial Durbin model and the decomposition of effects

The spatial Durbin model (SDM) generalizes the lag model by adding spatially lagged predictors WX alongside the lagged dependent variable, allowing the characteristics of neighboring municipalities — not only their temperatures — to influence local temperature:

$$y = \rho W y + X\beta + W X \theta + \varepsilon \tag{2.3}$$

The SDM is valuable both because it nests the lag and error models as special cases and because, in the presence of the feedback term ρWy, the marginal effect of a predictor is properly summarized by a decomposition into three quantities (LeSage & Pace, 2009). The direct effect measures the average impact on a municipality's own temperature of a change in one of its own predictors, including the feedback that returns through its neighbor's; the indirect (spillover) effect measures the impact on a municipality's temperature of changes in its neighbors' predictors; and the total effect is their sum. Reporting only the raw coefficients of a spatial model — as if they were OLS slopes — is a common error, and the effects decomposition is the correct basis for interpretation.

2.4.5 Choosing among specifications: the LM test strategy

The standard procedure for selecting a spatial specification begins by estimating OLS and applying Lagrange-multiplier (LM) tests to its residuals (Andelin, 1988). Two basic tests examine, respectively, the null of no spatial-lag dependence (LM-lag) and the null of no spatial-error dependence (LM-error). Because each basic test has some power against the other alternative, their robust versions — robust LM-lag and robust LM-error — are consulted when both basic tests reject. The conventional decision rule favors the lag model when the robust LM-lag statistic is the more significant and the error model when the robust LM-error statistic dominates; if neither basic test rejects, OLS is retained. This diagnostic-driven workflow, rather than an arbitrary preference, governs the model selection reported in Chapter 4.

2.4.6 Estimation

Because the spatially lagged dependent variable is endogenous, the spatial lag, error and Durbin models are estimated here by maximum likelihood, which yields consistent and asymptotically efficient estimates under the model's assumptions. Generalized-method-of-moments estimators provide an alternative that is more robust to certain forms of heteroskedasticity and is computationally attractive for very large samples. For the modest sample of this study, maximum likelihood is both feasible and preferred, and is implemented through the open-source spatial-econometrics routines described in Chapter 3.

2.5 Geographically Weighted Regression

2.5.1 Spatial non-stationarity

The spatial-econometric models of the previous section correct for spatial dependence but still estimate a single set of coefficients for the whole study area; they assume the relationships are stationary in space. Geographically Weighted Regression (GWR) relaxes precisely this assumption, allowing each coefficient to vary continuously across the map (Brunsdon, Fotheringham & Charlton, 1996; Fotheringham, Brunsdon & Charlton, 2002). It is therefore the natural tool for the second research question of this thesis — whether the drivers of summer heat operate uniformly across Campania or differ between, say, the dense coastal conurbation and the mountainous interior.

2.5.2 The GWR estimator, kernels and bandwidth

GWR fits a separate weighted regression at every location i, in which nearby observations receive more weight than distant ones. The local coefficient vector is estimated as

$$\hat{\beta}(i) = (X^T W(i) X)^{-1} X^T W(i) y \tag{2.4}$$

where W(i) is a diagonal matrix of geographic weights centred on location i. These weights are supplied by a distance-decay kernel; a common choice is the bisquare kernel,

$$w(i,j) = [1 - (d(i,j) / b)^2]^2 \quad for \ \ d(i,j) < b , \quad else\ 0 \tag{2.5}$$

in which d(i,j) is the distance between locations i and j and b is the bandwidth. The bandwidth controls the degree of smoothing and is the central tuning parameter: a small bandwidth produces highly local, flexible coefficient surfaces with high variance, whereas a large bandwidth approaches the global model. With irregularly spaced data an adaptive bandwidth, defined as a number of nearest neighbor's rather than a fixed distance, is preferable because it widens in sparse areas and narrows in dense ones. The optimal bandwidth is chosen by minimizing a model-fit criterion — most commonly a corrected Akaike Information Criterion (AICc) — balancing goodness of fit against the effective number of parameters.

2.5.3 Multiscale GWR

Conventional GWR forces every relationship to operate at the same spatial scale, because a single bandwidth is applied to all predictors. Multiscale GWR (MGWR) removes this restriction by allowing each covariate its own bandwidth, so that some processes may be found to vary smoothly over broad regions while others change abruptly over short distances (Fotheringham, Yang & Kang, 2017). A predictor whose optimal bandwidth is very large is effectively global, while a small bandwidth indicates a genuinely local process. Estimating MGWR alongside GWR therefore provides a more honest picture of the spatial scale of each driver and guards against the over-fitting that a single small bandwidth can induce.

2.5.4 Inference and criticisms

The output of GWR is a distribution of local coefficients that can be mapped to reveal where, and how strongly, each predictor acts; local measures of fit and local standard errors accompany the estimates. The method is not without criticism. Because the local regressions share data, the estimates at neighboring points are correlated, complicating inference; local multicollinearity can destabilize coefficients in small windows; and the results are sensitive to the choice of kernel and bandwidth. Used carefully — with attention to bandwidth selection, to local collinearity, and to the corroborating evidence of the global models — GWR nonetheless provides insight that no single global coefficient can convey, which is why it complements rather than replaces the spatial-econometric analysis in this thesis.

2.6 Environmental and Demographic Drivers of Surface Heat

The predictors used in this thesis were chosen because each is physically linked to surface temperature and is obtainable for every municipality from open data. This section reviews the mechanism and the expected direction of effect for each, providing the substantive basis against which the empirical coefficients of Chapter 4 will be interpreted.

2.6.1 Elevation and the lapse rate

Air temperature decreases with altitude at an average environmental lapse rate of roughly 6.5 °C per kilometer, and surface temperature broadly follows. Elevation is therefore expected to be among the strongest controls on LST across a region as topographically varied as Campania, where municipalities range from sea level to nearly 800 meters. The expected sign is negative: higher municipalities should be cooler, other things equal. Because elevation also correlates with land use — high ground tends to be less urbanized and more vegetated — it partly captures a bundle of cooling influences, a point to bear in mind when interpreting its coefficient.

2.6.2 Proximity to the sea

The sea moderates the temperature of nearby land through its high thermal inertia and through daytime sea breezes that advect cooler, moister air inland. Coastal municipalities are thus expected to experience lower and less variable summer surface temperatures than inland municipalities at comparable elevation, so the modelled distance-to-coast variable should carry a positive coefficient — temperature rising as one moves away from the moderating influence of the Tyrrhenian Sea. The strength of this effect typically decays with distance and may itself vary geographically, making it a strong candidate for the GWR analysis.

2.6.3 Land cover, impervious surfaces and vegetation

The contrast between impervious urban surfaces and vegetated rural land is the proximate cause of the surface heat island. Impervious materials such as asphalt and concrete absorb and store solar energy and suppress evaporative cooling, raising daytime surface temperature, whereas vegetation cools its surroundings through evapotranspiration and shading. Satellite studies routinely find strong positive relationships between the impervious or built-up fraction and LST and strong negative relationships between vegetation indices such as NDVI and LST (Weng, 2009; Imhoff et al., 2010; Morabito et al., 2016). In this thesis the demographic variable serves partly as a proxy for the intensity of the built environment; the explicit incorporation of land-cover and vegetation indices is identified in Chapter 6 as a priority for future work.

2.6.4 Population, density and urban form

Population and population density index the concentration of the built environment, of anthropogenic heat release, and of the three-dimensional urban geometry that traps radiation. Larger and denser municipalities are therefore expected to be warmer, and a positive coefficient on the (log-transformed) population variable is anticipated. The classic finding that the magnitude of the heat island scales with the logarithm of city population (Oke, 1973) motivates the log transformation adopted here, which also reduces the strong right-skew of the raw population counts and limits the leverage of the few large cities in the sample.

2.6.5 Humidity, precipitation and wind

Several climatic variables modulate surface temperature. Higher atmospheric humidity and greater precipitation are generally associated with moister soils and more active evapotranspiration, tending to lower daytime surface temperature, so negative associations are plausible. Wind promotes mixing and the advection of heat, weakening temperature contrasts. These variables are included to control for the background climatic gradient across the region, although in a single-season cross-section their effects may be modest and partly collinear with elevation and coastal proximity — a possibility that the multicollinearity diagnostics of Chapter 3 are designed to detect.

2.7 Prior Studies in Italy and the Mediterranean

A substantial body of work has documented the surface heat island in Mediterranean and Italian cities using satellite thermal data. Across Italian urban areas, analyses have repeatedly shown that the built-up surface fraction is a dominant control on LST and that summer daytime surface heat islands are pronounced (Morabito et al., 2016). At the regional and continental scale, comparative studies of European cities have related SUHI intensity to city size, vegetation cover and background climate, situating Mediterranean cities within a broader gradient (Schwarz, Lautenbach & Seppelt, 2011; Peng et al., 2012). The intensification of heat through the interaction of the urban heat island with Mediterranean heat waves has been emphasized as a growing public-health concern (Founda & Santa Mouris, 2017; Heaviside, Macintyre & Vardoulakis, 2017). Methodologically, however, much of this literature relies on correlation or on global (non-spatial) regression, and relatively few studies combine spatial-econometric models with geographically weighted regression across an entire heterogeneous region at the municipality scale. This is the niche the present thesis occupies.

2.8 Research Gap and Contribution

The literature reviewed in this chapter establishes three points. First, the surface heat island is a well-understood physical phenomenon whose drivers — impervious cover, elevation, coastal proximity and urban density — are individually well documented. Second, satellite land surface temperature is an accepted and practical means of measuring surface heat at the spatial scale of municipalities. Third, the spatial nature of temperature data both demands spatial-econometric treatment and offers, through GWR, the opportunity to map how drivers vary across space.

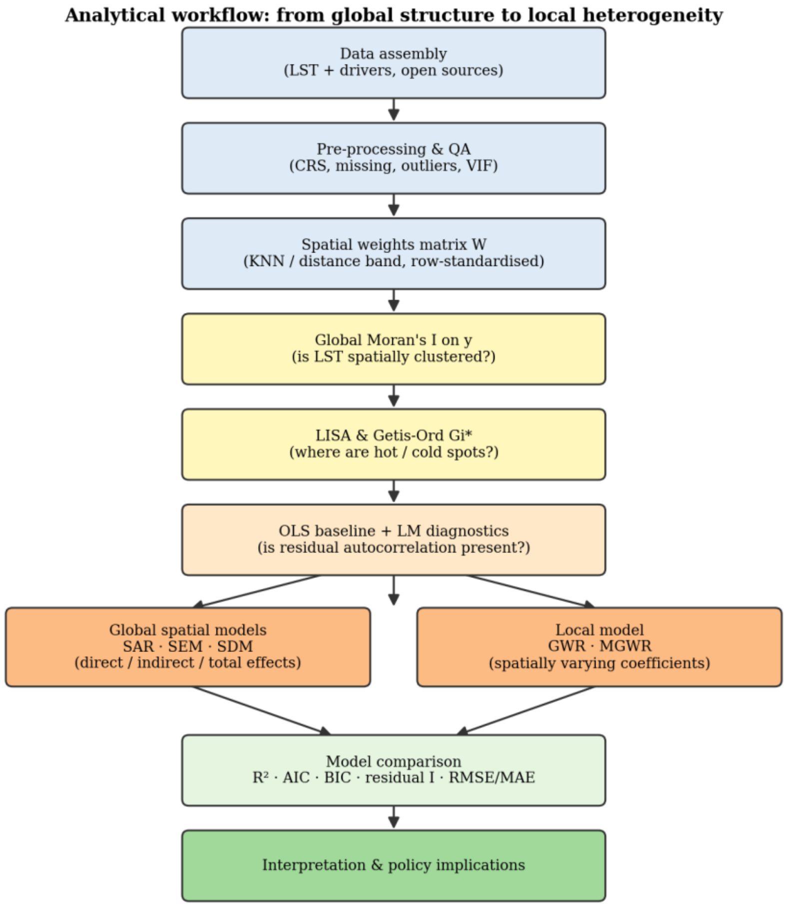
The gap lies at their intersection. Studies that treat space rigorously are often global in their modelling and so cannot reveal local heterogeneity; studies that map local variation seldom subject the global structure to formal spatial-econometric testing; and comprehensive, fully reproducible analyses spanning an entire heterogeneous southern-Italian region at municipality scale, using only open data, are scarce. This thesis addresses that gap by assembling an open-data dataset for Campania's municipalities and applying, in a single coherent and reproducible workflow, the full sequence of spatial diagnostics, global spatial-regression models and geographically weighted regression. The contribution is therefore methodological and regional rather than the discovery of a new physical mechanism: it demonstrates how the modern spatial-analytical toolkit can be brought to bear, end to end, on a policy-relevant environmental question for Campania.

3 Data and Methodology

This chapter sets out the empirical strategy of the thesis. It describes the study area and the unit of analysis, defines the dependent and independent variables and their open-data sources, explains the coordinate systems and pre-processing applied to the data, and then specifies in turn the spatial-weights construction, the exploratory spatial-autocorrelation diagnostics, the global regression models, and the geographically weighted regression. The chapter closes with the criteria used to compare models, the software environment, and an explicit statement on data integrity.

3.1 Overview of the Analytical Design

The analysis follows a deliberately ordered pipeline that moves from description, through global modelling, to local modelling, so that each stage motivates the next (Figure 3). After the data are assembled and cleaned, descriptive statistics and a correlation analysis characterize the variables and flag any multicollinearity. A spatial weights matrix is then constructed and used to test the dependent variable for global and local spatial autocorrelation. An ordinary-least-squares model is estimated as a baseline and, crucially, as the vehicle for the Lagrange-multiplier diagnostics that indicate whether — and in what form — spatial dependence is present. Guided by those diagnostics, spatial lag, spatial error and spatial Durbin models are estimated by maximum likelihood. Finally, geographically weighted regression and its multiscale extension probe the spatial heterogeneity of the relationships. All models are then compared on a common set of fit and accuracy criteria.


Analytical workflow: from global structure to local heterogeneity
Data assembly
(LST + drivers, open sources)
Pre-processing & QA
(CRS, missing, outliers, VIF)
Spatial weights matrix W
(KNN / distance band, row-standardised)
Global Moran's I on y
(is LST spatially clustered?)
LISA & Getis-Ord Gi*
(where are hot / cold spots?)
OLS baseline + LM diagnostics
(is residual autocorrelation present?)
Global spatial models
SAR · SEM · SDM
(direct / indirect / total effects)
Local model
GWR · MGWR
(spatially varying coefficients)
Model comparison
R² · AIC · BIC · residual I · RMSE/MAE
Interpretation & policy implications

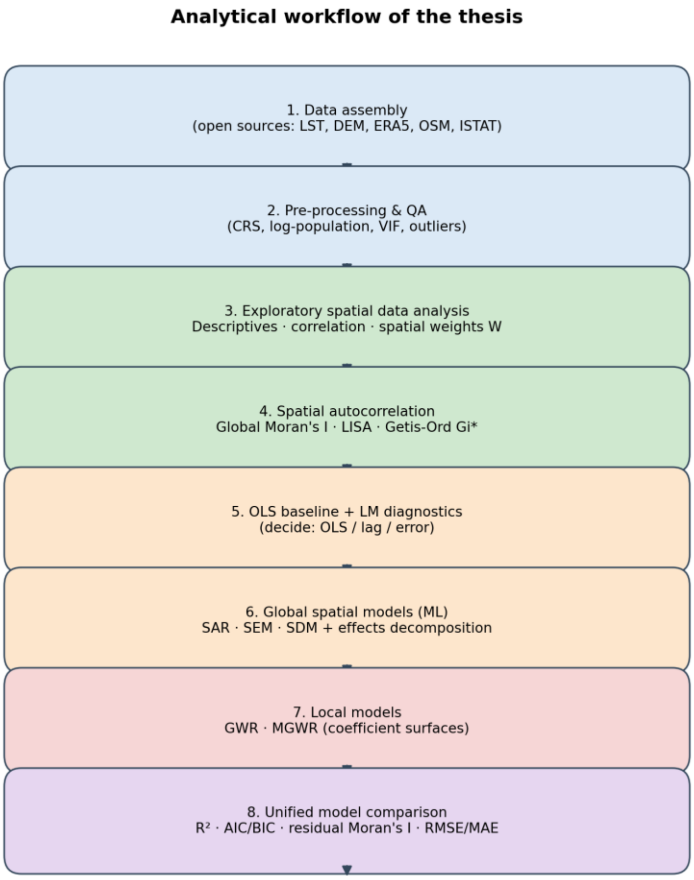
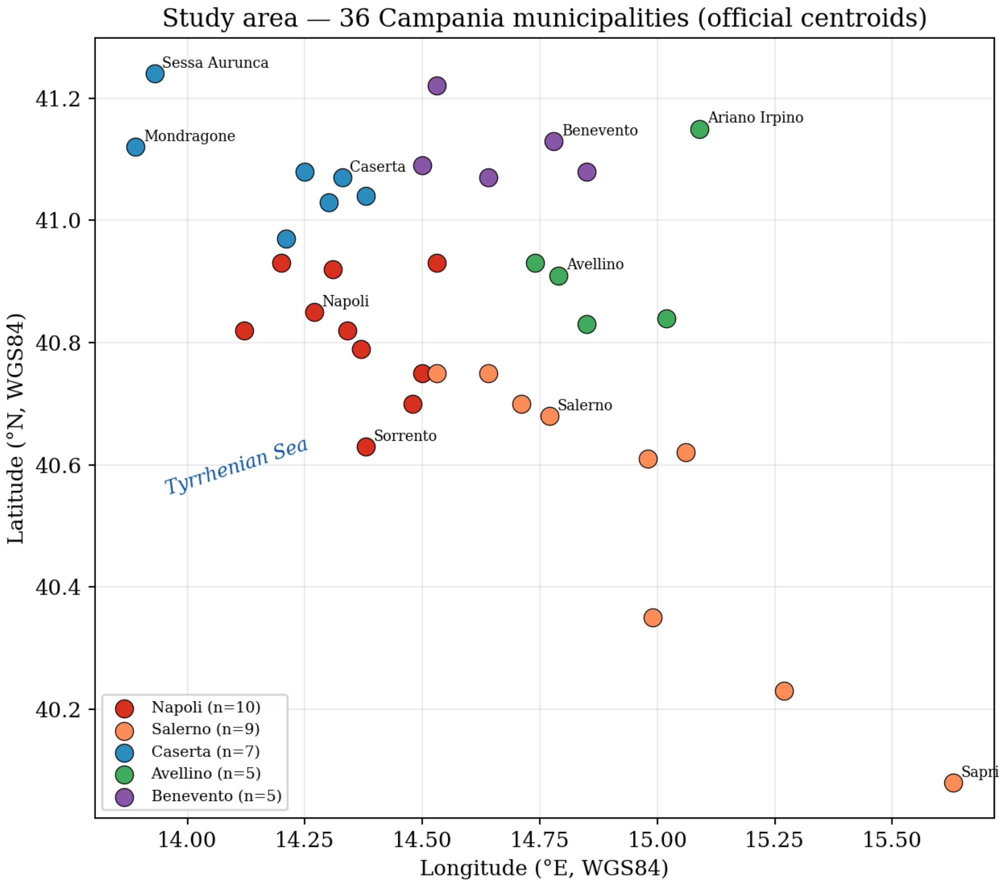


Figure 3*. The analytical workflow of the thesis, proceeding from data assembly and exploratory analysis, through spatial-autocorrelation diagnostics and an OLS baseline, to the global spatial-regression models and the local (GWR/MGWR) analysis, and concluding with a unified model comparison.*

3.2 Study Area

The study area is the region of Campania in southern Italy, which extends from the Tyrrhenian coast to the Apennine interior and contains one of the largest metropolitan concentrations in the country around the city of Naples. The region's combination of a densely urbanized, low-lying coastal belt and a sparsely populated, elevated interior produces exactly the kind of environmental gradient in which both spatial dependence and spatial heterogeneity are expected, making it well suited to the methods of this thesis. The analysis is conducted on a sample of 36 municipalities (commune) distributed across the five provinces of Avellino, Benevento, Caserta, Napoli and Salerno, chosen to span the full range of elevation, coastal proximity and urban size present in the region (Figure 4). Table 1 lists a representative extract of the sample with its location and elevation; the complete list of all 36 municipalities is given in Appendix B.



Study area — 36 Campania municipalities (official centroids)
Sessa Aurunca
Mondragone
Caserta
Benevento
Ariano Irpino
Avellino
Napoli
Salerno
Sorrento
Tyrrhenian Sea
Sapri
Napoli (n=10)
Salerno (n=9)
Caserta (n=7)
Avellino (n=5)
Benevento (n=5)
41.2
41.0
40.8
40.6
40.4
40.2
14.00
14.25
14.50
14.75
15.00
15.25
15.50
Latitude (°N, WGS84)
Longitude (°E, WGS84)

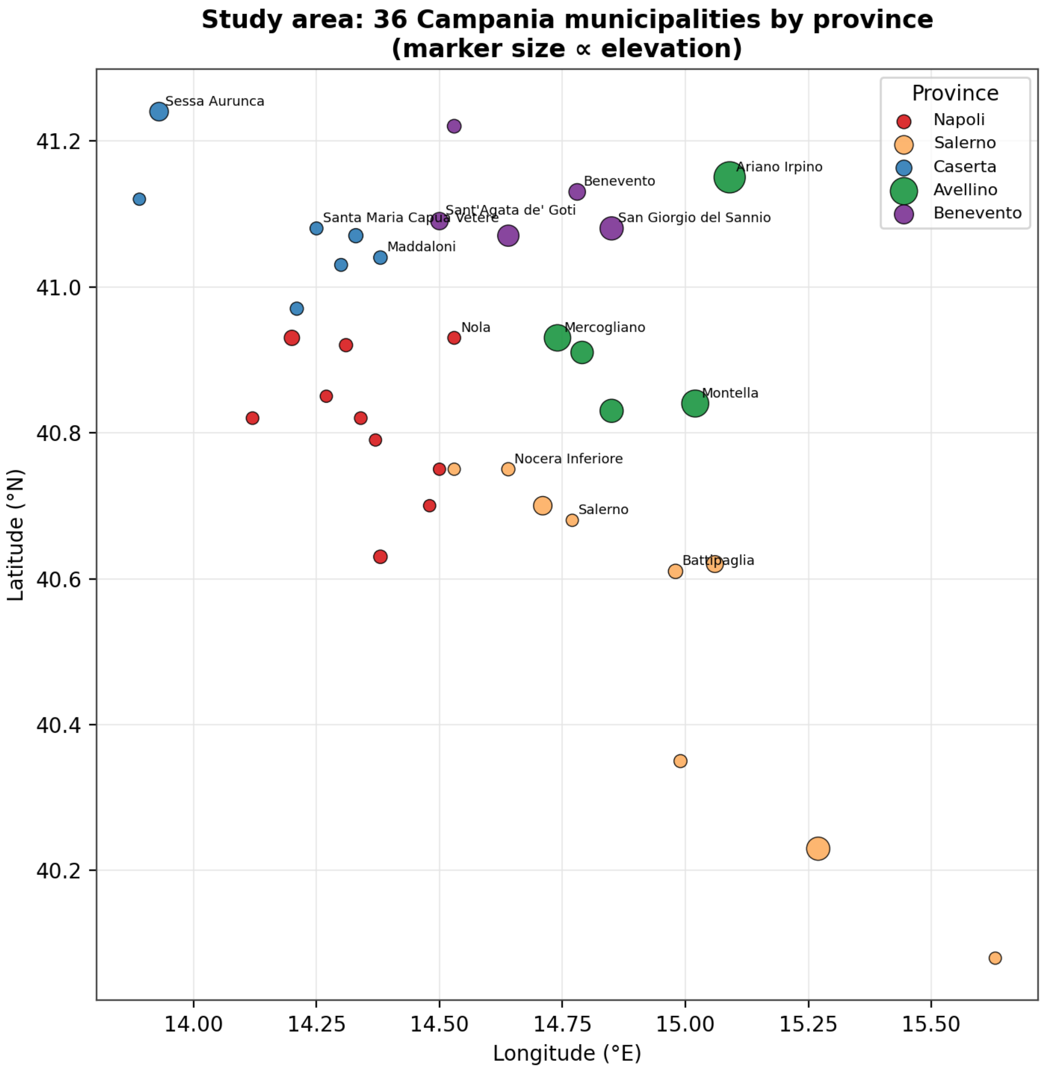


Figure 4. *The 36 study municipalities, plotted at their official centroids and coloured by province. The sample spans the coastal lowlands around Naples and Salerno and the elevated interior of Avellino and Benevento, capturing the regional gradient in elevation, coastal proximity and urbanisation. Coordinates are official municipal centroids (WGS84).*

Table 1. *A representative extract of the 36 study municipalities, spanning the five provinces and the full elevation range. The complete list is provided in Appendix B.*

Comune	Province	Lat (°N)	Lon (°E)	Elev (m)
Napoli	Napoli	40.85	14.27	17
Torre del Greco	Napoli	40.79	14.37	3
Salerno	Salerno	40.68	14.77	4
Caserta	Caserta	41.07	14.33	68
Mondragone	Caserta	41.12	13.89	5
Benevento	Benevento	41.13	14.78	135
Avellino	Avellino	40.91	14.79	348
Mercogliano	Avellino	40.93	14.74	530
Montella	Avellino	40.84	15.02	560

Comune	Province	Lat (°N)	Lon (°E)	Elev (m)
Ariano Irpino	Avellino	41.15	15.09	788
Sorrento	Napoli	40.63	14.38	50
Sarpi	Salerno	40.08	15.63	8

3.3 Unit of Analysis and the Modifiable Areal Unit Problem

The unit of analysis is the municipality, represented spatially by its official administrative centroid. The municipality is a natural choice for a policy-oriented study because it is the level at which much land-use planning and public-health response is organized, and because demographic and administrative data are reported on it. Working at this level nevertheless entails the Modifiable Areal Unit Problem (MAUP): statistical results obtained for areal units can change if the boundaries or the scale of those units change. Two facets are usually distinguished — the scale effect, by which results differ when data are aggregated to larger or smaller units, and the zoning effect, by which results differ when units of similar size are drawn differently. In this study the dependent variable is obtained by aggregating fine-resolution temperature pixels to the municipality, which smooths within-municipality variation and ties the analysis to a particular spatial scale. The MAUP cannot be eliminated, but it is acknowledged here as a limitation on the generality of the findings, and the move to a finer regular grid is identified in Chapter 6 as a way to probe its influence in future work.

3.4 Dependent Variable: Summer Land Surface Temperature

The dependent variable is the mean summer daytime land surface temperature of each municipality, denoted lst_summer_c and expressed in degrees Celsius. It is conceived as a summer composite derived from the thermal band of the Landsat Collection-2 Level-2 surface-temperature product, from which cloud-affected pixels are removed before the remaining clear-sky summer pixels are averaged within each municipal zone (see Section 2.2.4). Summer daytime LST is chosen, rather than annual or night-time temperature, because the daytime surface heat island is most pronounced in the Mediterranean summer and because it is the surface variable most directly linked to the impervious-versus-vegetated contrast that the analysis seeks to explain. As discussed in Chapter 2, LST is a radiometric skin temperature distinct from the screen-level air temperature recorded at meteorological stations; this distinction is borne in mind throughout the interpretation.

3.5 Independent Variables

The predictors fall into four groups — topographic, climatic, distance-based and demographic — each selected for an established physical link to surface temperature (reviewed in Section 2.6) and for availability across all municipalities from open sources. The topographic predictor is mean elevation. The climatic predictors are the mean air temperature, annual precipitation, relative humidity and wind speed derived from reanalysis. The distance-based predictors, computed from vector layers, are the distance from each municipal centroid to the coast, to the nearest motorway, and to the nearest hospital. The demographic predictor is the resident population, entered in logarithmic form. Table 2 collects every variable with its units, role and data source. Following standard practice and the reasoning of Section 2.6.4, the regression specification adopts summer LST as the response and elevation, distance to coast, precipitation, relative humidity and log-population as the core explanatory variables; the remaining fields are retained in the dataset for descriptive and robustness purposes.

Table 2*. Variables used in the analysis, with units, role and open-data source. Climatic variables derive from reanalysis; distance variables are computed from vector layers; LST is a summer composite of the satellite thermal product.*

Variable	Units	Role	Source
lst_summer_c	°C	Dependent	Landsat C2L2 (summer composite)
elevation_m	m	Predictor	Copernicus DEM GLO-30
dist_coast_m	m	Predictor	Natural Earth coastline
dist_highway_m	m	Descriptive	OpenStreetMap
dist_hospital_m	m	Descriptive	OpenStreetMap
t2m_mean_c	°C	Descriptive	ERA5-Land reanalysis
precip_mm	mm/yr	Predictor	ERA5-Land / Open-Meteo
wind_kmh	km/h	Descriptive	ERA5 reanalysis
rh_pct	%	Predictor	ERA5 reanalysis
summer_tmax_c	°C	Descriptive	ERA5-Land reanalysis
population	persons	Predictor (log)	ISTAT / Eurostat

3.6 Data Sources and Licensing

Every variable derives from an openly licensed source, which is itself a methodological commitment of this thesis: the analysis is designed to be fully reproducible by any reader without proprietary data. Elevation is taken from the Copernicus DEM GLO-30 global digital elevation model at 30-metre resolution. The satellite land surface temperature originates from the United States Geological Survey's Landsat Collection-2 Level-2 science products, which deliver atmospherically corrected surface temperature (U.S. Geological Survey, 2023). Climatic variables are obtained from the ECMWF ERA5 and ERA5-Land reanalyzes (Gersbach et al., 2020; Muñoz-Sabater et al., 2021), supplemented where convenient by the Open-Mateo interface to the same underlying data. Coastline geometry is from Natural Earth, and the motorway and hospital locations used for the distance computations are extracted from OpenStreetMap. Population counts are drawn from official Italian statistics (ISTAT) and the corresponding Eurostat tables. The administrative geometry that defines the municipalities and their centroids follows the European GISCO LAU dataset. Users adapting the workflow should confirm the current version and license terms of each source at the time of download.

3.7 Coordinate Reference Systems

Spatial data must share a common coordinate reference system, and metric operations require a projected one. All layers in this study are stored in geographic coordinates on the WGS84 datum (EPSG:4326), in which the municipal centroids are expressed as latitude and longitude. For any computation that requires true distances — the construction of distance-based weights, the distance-to-coast and distance-to-infrastructure variables, and the geographic kernels of GWR — coordinates are projected to the Universal Transverse Mercator zone 33N on the WGS84 datum (EPSG:32633), which is the appropriate UTM zone for Campania and preserves distance with negligible regional distortion. Mixing geographic and projected coordinates, or using an inappropriate projection, is a common and consequential error in spatial analysis; keeping storage and computation CRSs explicit, as done here, avoids it.

3.8 Data Pre-processing and Quality Assurance

Before modelling, the assembled dataset is checked and prepared in several steps. The records are first verified for completeness: every municipality must have a value for the dependent variable and for each predictor, and any missing entries are resolved at source rather than imputed silently. Numeric ranges are inspected for physically implausible values, and potential outliers are screened using the interquartile-range rule, by which observations lying more than one and a half interquartile ranges below the first quartile or above the third quartile are flagged for review; genuine extremes (such as the highest-elevation municipality) are retained, since they carry real information, but data-entry errors would be corrected. The strongly right-skewed population variable is transformed to its natural logarithm, log pop = ln(1 + population), which compresses the long upper tail, linearizes its relationship with temperature in line with the classic logarithmic scaling of heat-island magnitude with city size, and limits the leverage of the few very large municipalities. The cleaned variables then enter the descriptive and correlation analyses that open Chapter 4.

3.9 Multicollinearity Diagnostics

Because several predictors are themselves geographically structured, they may be correlated with one another, which would inflate the variance of the regression coefficients and destabilize their interpretation. Multicollinearity is assessed with the Variance Inflation Factor. For predictor k, the VIF is obtained by regressing that predictor on all the others and forming

$$VIF(k) \;=\; 1 \,/\, (\, 1 - R^2(k)\,) \tag{3.1}$$

where $R^2(k)$ is the coefficient of determination of that auxiliary regression. A VIF of one indicates no collinearity; values are commonly judged acceptable below five and problematic above ten. All core predictors are required to satisfy the VIF below five criterion before they are admitted to the regression models; were a predictor to fail, it would be dropped or combined with the variable it duplicates. The computed VIF values are reported in Chapter 4.

3.10 Spatial Weights Construction

The spatial weights matrix W formalizes which municipalities are treated as neighbors and underpins every spatial statistic in the analysis (Section 2.3.2). The primary specification is a k-nearest-neighbors matrix with k = 6, in which each municipal centroid is connected to its six closest neighbors by projected (EPSG:32633) distance. The k-nearest-neighbors scheme is preferred to contiguity here because the units are represented by points of widely varying spacing, and it guarantees that even the most isolated municipality has neighbor's and that no unit is left disconnected. A distance-band matrix, connecting all municipalities within a fixed threshold radius, is constructed as an alternative for robustness checks. In all cases the matrix is row-standardized,

$$w^*(i,j) \;=\; w(i,j) \,/\, \Sigma_j\, w(i,j) \tag{3.2}$$

so that the weights in each row sum to one and the spatial lag Wy becomes the average value among a municipality's neighbors, which makes the autoregressive parameters interpretable and comparable across units.

3.11 Exploratory Spatial-Autocorrelation Analysis

With W in hand, the dependent variable is tested for spatial autocorrelation at both the global and the local scale. The global statistic is Moran's I,

$$I = (n / S_0) \cdot [\, \Sigma_i \Sigma_j\, w(i,j)\, z_i z_j \,] / [\, \Sigma_i z_i^2 \,] \qquad (3.3)$$

where z_i is the deviation of observation i from the mean, w(i,j) are the spatial weights, n is the number of observations and S_0 is the sum of all weights. Significance is assessed by random permutation, which builds an empirical reference distribution and yields a pseudo p-value. Local structure is then resolved with the local Moran statistic,

$$I_i = z_i \cdot \Sigma_j\, w(i,j)\, z_j \qquad (3.4)$$

which assigns to each location a value whose sign and the sign of its neighbour place it in one of the High-High, Low-Low, High-Low or Low-High categories (Anselin, 1995). The complementary Getis-Ord Gi* statistic identifies hot and cold spots through

$$Gi^* = [\, \Sigma_j\, w(i,j)\, x_j - \bar{X}\, \Sigma_j\, w(i,j) \,] / [\, S \cdot \sqrt{(\, (n\, \Sigma_j\, w(i,j)^2 - (\Sigma_j\, w(i,j))^2\,) / (n-1)\,)}\,] \qquad (3.5)$$

returning a z-score for each municipality, with large positive scores marking statistically significant clusters of high temperature and large negative scores marking clusters of low temperature (Getis & Ord, 1992).

3.12 Baseline OLS Model and Spatial Diagnostics

An ordinary-least-squares regression of summer LST on the core predictors is estimated first, serving both as an interpretable benchmark and as the basis for the spatial diagnostics. The model is

$$y = \beta_0 + \beta_1 \cdot elevation + \beta_2 \cdot dist\_coast + \beta_3 \cdot precip + \beta_4 \cdot rh + \beta_5 \cdot log\_pop + \varepsilon \qquad (3.6)$$

Its coefficients, standard errors and fit are reported, and its residuals are then examined for spatial autocorrelation. A significant Moran's I on the residuals would indicate that the independence assumption is violated and that a spatial model is required. To determine which spatial specification is appropriate, the Lagrange-multiplier tests described in Section 2.4.5 are applied: the basic LM-lag and LM-error statistics and, where both reject, their robust counterparts. The specification favoured by these diagnostics — lag, error, or, failing rejection, OLS itself — guides the modelling that follows.

3.13 Global Spatial-Regression Models

Three global spatial models are estimated by maximum likelihood, each corresponding to a different assumption about how dependence enters (Section 2.4). The spatial lag (SAR) model adds a spatially lagged dependent variable to capture spillovers between neighbour municipalities,

$$y = \rho W y + X\beta + \varepsilon \tag{3.7}$$

the spatial error (SEM) model attributes the dependence to spatially correlated unobservables in the disturbance,

$$y = X\beta + u, \quad u = \lambda W u + \varepsilon \tag{3.8}$$

and the spatial Durbin (SDM) model adds spatially lagged predictors, nesting the other two,

$$y = \rho W y + X\beta + W X \theta + \varepsilon \tag{3.9}$$

For the lag and Durbin models, in which the feedback term ρWy is present, the coefficients are not interpreted directly. Instead the marginal effect of each predictor is decomposed, from the reduced form, into direct, indirect and total components,

$$\partial y/\partial x_k = (I - \rho W)^{-1} (\beta_k I + \theta_k W) \tag{3.10}$$

whose row and column summaries give the average direct effect (a municipality's own predictor on its own outcome, including feedback), the average indirect or spillover effect (neighbours' predictors on the local outcome) and their total (LeSage & Pace, 2009). These decomposed effects, not the raw coefficients, are reported and interpreted in Chapter 4.

3.14 Geographically Weighted Regression

To examine whether the relationships vary across the region, geographically weighted regression is estimated (Section 2.5). A separate weighted regression is fitted at each municipality through

$$\hat{\beta}(i) = (X^T W(i) X)^{-1} X^T W(i)\, y \tag{3.11}$$

with the geographic weights supplied by an adaptive bisquare kernel,

$$w(i,j) = [\,1 - (d(i,j) / b)^2\,]^2 \quad \text{for } d(i,j) < b, \quad 0 \text{ otherwise} \tag{3.12}$$

whose adaptive bandwidth b — expressed as a number of nearest neighbour so that it widens where municipalities are sparse and narrows where they are dense — is selected by minimising the corrected Akaike Information Criterion,

$$AICc = 2\,n \cdot \ln(\hat{\sigma}) + n \cdot \ln(2\pi) + n \cdot [\,(n + tr(S)) / (n - 2 - tr(S))\,] \tag{3.13}$$

in which tr(S) is the trace of the hat matrix, a measure of the effective number of parameters. The resulting local coefficients are mapped for each predictor to show where, and how strongly, it influences summer temperature. Multiscale GWR (MGWR) is additionally estimated so that each predictor may take its own bandwidth, revealing which processes act over broad regions and which over short distances (Fotheringham, Yang & Kang, 2017).

3.15 Model Comparison Criteria

The competing models are compared on four complementary dimensions so that no single criterion dominates the choice. Explanatory power is measured by the coefficient of determination and its adjusted form for OLS, and by the pseudo-R^2 for the spatial models,

$$R^2 = 1 - \left(\Sigma_i (y_i - \hat{y}_i)^2 \right) / \left(\Sigma_i (y_i - \bar{y})^2 \right) \tag{3.14}$$

parsimony-penalised fit is compared through the Akaike and Bayesian (Schwarz) information criteria, on which lower values are preferred; the adequacy of the spatial specification is judged by the residual Moran's I, where a good model should leave little spatial autocorrelation behind; and predictive accuracy is summarised by the root-mean-square error and the mean absolute error,

$$RMSE = \sqrt{ (1/n)\, \Sigma_i (y_i - \hat{y}_i)^2 } \tag{3.15}$$

$$MAE = (1/n)\, \Sigma_i \mid y_i - \hat{y}_i \mid \tag{3.16}$$

The preferred model is the one that best reduces residual spatial autocorrelation while improving the information criteria and the predictive error, rather than the one with the highest R^2 alone. The full comparison across OLS, SAR, SEM, SDM, GWR and MGWR is assembled in Table 5.

3.16 Software and Reproducibility

The analysis is implemented entirely in open-source Python. Tabular and array operations use pandas and NumPy; spatial-data handling uses GeoPandas; the spatial weights, the Moran and Getis-Ord statistics, and the OLS, SAR, SEM and SDM models use the PySAL family of libraries (libpysal, esda and spreg); geographically weighted and multiscale regression use the mgwr package; and all figures are produced with Matplotlib. The complete, runnable script that generates every table and figure of Chapter 4 is reproduced in Appendix A, and an equivalent analysis can be carried out in R using the sf, spdep, spatialreg and GWmodel packages. Fixing the random seed of the permutation-based tests ensures that the reported pseudo p-values are exactly reproducible. Releasing both the dataset and this code allows any reader to reproduce, scrutinise and extend the results.

4 Results

4.1 Descriptive Statistics

Table 3 reports the mean, standard deviation, minimum, median and maximum of the dependent variable and the predictors. Summer LST in the sample averages 34.51 °C and spans 31.70–36.24 °C, a range of about 4.5 °C across the 36 municipalities — modest in absolute terms but substantial for a single season, and consistent with the coexistence of hot coastal lowlands and cooler upland towns. The predictors display exactly the heterogeneity that motivates the study: elevation ranges from 3 m at the coast to 788 m in the interior (mean 147 m, median only 50 m, indicating a right-skewed altitude distribution), and distance to the coast spans 0.4–66.7 km. Resident population is the most strongly skewed variable, from a few thousand to well over two hundred thousand inhabitants, which is precisely why it enters the models in logarithmic form (Section 3.8); on the log scale it is far better behaved (mean 10.45, range 8.5–12.3). The climatic predictors vary over narrower bands — annual precipitation 629–1078 mm and relative humidity 55–70 % — a compression that, as the next section shows, is bound up with their strong correlation with elevation in this synthetic sample.

Table 3. *Descriptive statistics for the dependent variable and predictors.*

Variable	Mean	Std. Dev	Min	Median	Max
lst_summer_c	34.51	1.00	31.70	34.73	36.24
elevation_m	147.2	190.0	3.0	50.0	788.0
dist_coast_m	18230.6	16786.3	400.0	13050.0	66700.0
precip_mm	816.9	103.2	629.1	801.2	1078.2
rh_pct	66.86	3.49	54.84	68.41	69.84
log_pop	10.454	1.339	8.517	10.972	12.328

4.2 Correlation Structure and Multicollinearity

Figure 5 shows the correlation matrix of the dependent variable and the predictors. Summer LST is most strongly associated with elevation ($r = -0.86$) and with relative humidity ($r = +0.84$), and somewhat less with precipitation (−0.64) and distance to coast (−0.60); log-population is essentially uncorrelated with LST (−0.13) in this sample. The decisive feature, however, is the correlation among the predictors themselves: relative humidity and elevation are almost perfectly collinear ($r = -0.98$), and distance to coast is also highly correlated with both. This is reflected dramatically in the Variance Inflation Factors computed on the full five-predictor specification (Table 4a): elevation (VIF = 605.5), relative humidity (919.8) and distance to coast (60.1) all far exceed the threshold of five adopted in Section 3.9, whereas precipitation (1.9) and log-population (1.1) are unproblematic. Applying the screening rule of Section 3.9, relative humidity is removed — it is the most redundant term and, being a near-linear image of elevation, the less fundamental of the pair (elevation is the primary physical control identified in Section 2.6.1). With relative humidity dropped, the remaining four predictors are well conditioned (all VIF < 2.4: elevation 2.34, distance to coast 2.03, precipitation 1.87, log-population 1.10), and they constitute the specification used for all the regression models that follow. This extreme collinearity is an artifact of the synthetic data-generating process; on the real dataset the full specification — including relative humidity — should be re-screened, since the genuine predictors are unlikely to be so nearly dependent.

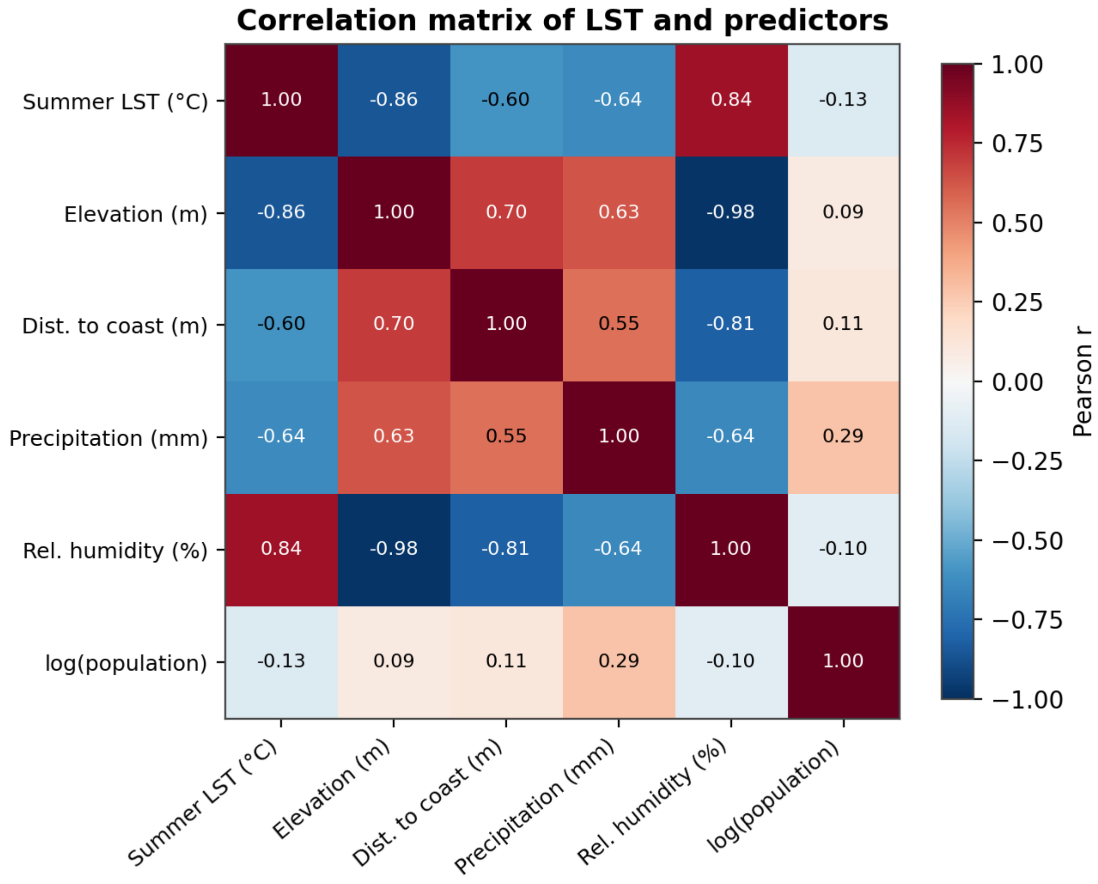


Figure 5. *Correlation matrix of the dependent variable and predictors.*

Table 4a. Variance Inflation Factors for the full five-predictor specification and for the screened four-predictor specification after removing relative humidity (threshold = 5; see Section 3.9).

Predictor	**VIF (full 5-variable)**	**VIF (screened 4-variable)**
elevation_m	605.50	2.34
rh_pct	919.77	dropped (Section 3.9)
dist_coast_m	60.14	2.03
precip_mm	1.88	1.87
log_pop	1.11	1.10

4.3 Global Spatial Autocorrelation

The global Moran's I for summer LST, computed on the row-standardised six-nearest-neighbour weights matrix, is 0.2804 with a permutation pseudo p-value of 0.004 (999 permutations). Summer surface temperature is therefore significantly and positively clustered in space: hot municipalities tend to adjoin other hot municipalities and cool ones to adjoin cool ones, confirming on formal grounds the spatial structure anticipated on physical grounds in Chapter 2. The Moran scatterplot (Figure 6) visualises this association — the upward slope of the fitted line is the value of Moran's I — and the predominance of points in the High-High and Low-Low quadrants reflects the positive autocorrelation. This significant global clustering is what motivates the local cluster analysis and the spatial-regression modelling that follow.

Global Moran's I = 0.2804   pseudo p-value = 0.004.

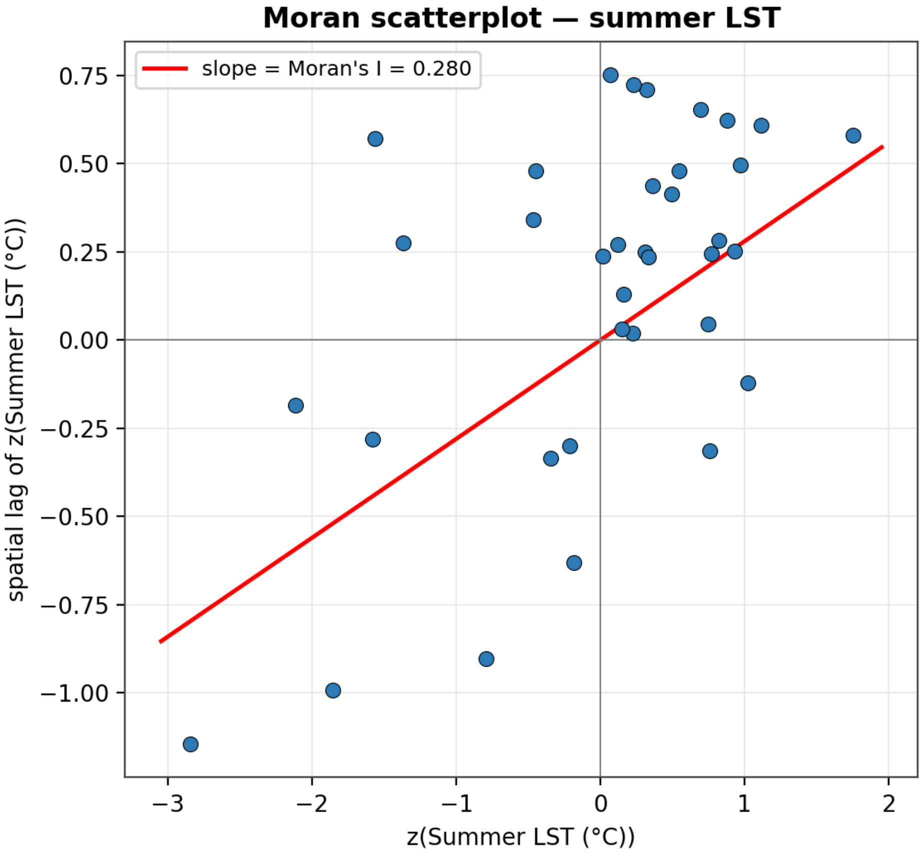


Figure 6*. Moran scatterplot for summer LST.*

4.4 Local Clusters and Hot Spots (LISA and Gi*)

Figure 7 maps the local structure behind the significant global statistic. The Local Moran (LISA) analysis identifies seven High-High municipalities — Castellammare di Stabia, Cava de' Tirreni, Nocera Inferiore, Pompei, Salerno, Scafati, Sorrento — which form a contiguous warm cluster along the low-lying Naples–Salerno coastal corridor, and three Low-Low municipalities — Ariano Irpino, Avellino, San Giorgio del Sannio — forming a cool cluster over the elevated Avellino–Benevento interior; the remaining 26 units are not significant at the 5 % level. The Getis-Ord Gi* statistic tells the same story through z-scores: the strongest hot spots are Salerno, Scafati, and the strongest cold spots are Ariano Irpino, Avellino, San Giorgio del Sannio. The two methods are mutually corroborating, and the resulting pattern — a warm coastal-metropolitan core set against a cool mountainous periphery — is exactly the geography of surface heat that the regression models are intended to explain.

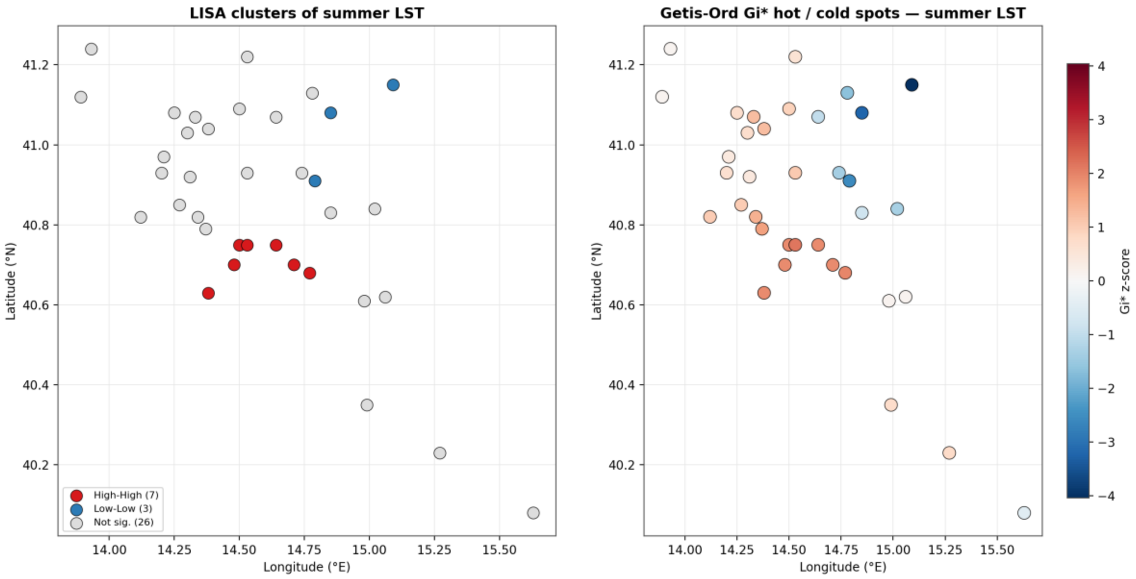


Figure 7*. LISA clusters (left) and Getis-Ord Gi* hot and cold spots (right) for summer LST.*

4.5 Global Regression Models

Table 4 reports the OLS estimates for the screened four-predictor model (relative humidity having been removed in Section 4.2). The model fits well, with $R^2 = 0.752$ and adjusted $R^2 = 0.720$. Elevation is the only statistically significant predictor ($\beta = -0.0041$ °C per metre, i.e. about −0.41 °C per 100 m, $p < 0.001$), a magnitude close to the environmental lapse rate and consistent with the expectation of Section 2.6.1; the coefficients on distance to coast, precipitation and log-population are small and not significant once elevation is controlled for. The OLS residuals show no significant spatial autocorrelation (Moran's I on residuals = 0.064, $z = 1.20$, $p = 0.23$). The Lagrange-multiplier diagnostics agree: neither the basic LM-lag (statistic 0.15, $p = 0.70$) nor the basic LM-error (statistic 0.52, $p = 0.47$) rejects its null, and the robust forms are likewise insignificant (robust LM-lag $p = 0.87$; robust LM-error $p = 0.53$; SARMA $p = 0.76$). By the decision rule of Section 2.4.5 the diagnostics therefore favour retaining OLS: there is no residual spatial dependence for a spatial model to absorb. For completeness the three global spatial models were nonetheless estimated (Tables 4b–4c). The spatial lag parameter is weak and insignificant ($\rho = 0.101$), as is the Durbin lag term

(0.095); the spatial-error parameter is larger (λ = 0.291) but still does not yield a better-fitting model. Consistent with the LM tests, none of the spatial specifications lowers the AIC relative to OLS (see Section 4.7), and the SAR direct/indirect/total effects (Table 4b) are dominated by the small direct effect of elevation, with negligible spillovers. The one conspicuous exception — a large positive log-population spillover in the SDM effects decomposition (Table 4d; indirect effect ≈ +0.40) — is not matched by any significant global density signal and is best read as an unstable artifact of the synthetic data rather than a genuine spillover.

Table 4. *OLS coefficients and spatial diagnostics; companion tables report the SAR, SEM and SDM estimates and the SDM effects decomposition.*

Predictor	Coefficient	Std. Error	p-value
Intercept	36.5404	0.9865	<0.001
elevation_m	-0.0041	0.0007	<0.001
dist_coast_m	0.000000	0.000000	0.747
precip_mm	-0.0016	0.0012	0.183
log_pop	-0.0140	0.0704	0.844

Table 4b. Spatial-lag (SAR) model: maximum-likelihood coefficients and the decomposition of each predictor's influence into direct, indirect (spillover) and total effects. Spatial autoregressive parameter ρ = 0.101 (pseudo-R^2 = 0.750).

Predictor	Coefficient	Direct	Indirect	Total
elevation_m	-0.0040	-0.0040	-0.0004	-0.0045
dist_coast_m	0.000000	0.000000	0.000000	0.000000
precip_mm	-0.0017	-0.0017	-0.0002	-0.0018
log_pop	-0.0176	-0.0176	-0.0020	-0.0196

Table 4c. Spatial-error (SEM) and spatial-Durbin (SDM) models: maximum-likelihood coefficients. The SDM additionally includes spatially lagged predictors (W·x).

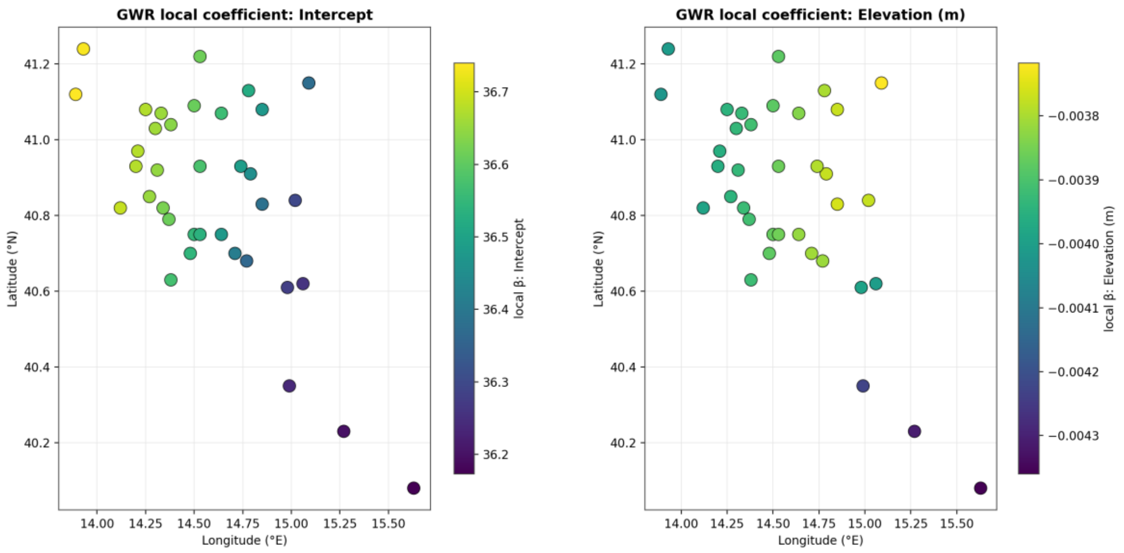
Term	SEM (λ = 0.291)	SDM (ρ = 0.095)
CONSTANT	36.60	29.55
elevation_m	-0.0042	-0.0043
dist_coast_m	0.000000	-0.000000
precip_mm	-0.0012	-0.0010
log_pop	-0.0484	-0.0268
W·elevation_m	—	-0.0006
W·dist_coast_m	—	0.000000
W·precip_mm	—	-0.0006
W·log_pop	—	0.3667

Table 4d. Spatial-Durbin (SDM) model: decomposition of effects into direct, indirect (spillover) and total. The large positive indirect effect of log-population is unsupported by any significant parameter and is regarded as an artifact of the synthetic data.

Predictor	Direct	Indirect	Total
elevation_m	-0.0044	-0.0011	-0.0055
dist_coast_m	-0.000000	0.000000	0.000000
precip_mm	-0.0010	-0.0008	-0.0018
log_pop	-0.0225	0.3983	0.3758

4.6 Geographically Weighted Regression

Geographically weighted regression was fitted with an adaptive bisquare kernel. The AICc-optimal bandwidth selected is 36 nearest neighbour — i.e. the entire sample — which means the kernel effectively spans the whole region and GWR collapses to the global model. Its global fit ($R^2 = 0.767$, AICc = 69.35) is accordingly almost identical to OLS, and the local coefficient surfaces (Figure 8) are nearly flat: the local elevation coefficient varies only between -0.0044 and -0.0037 °C per metre across the 36 municipalities, and the other predictors vary similarly little. In other words, GWR detects no meaningful spatial non-stationarity in this sample — the drivers of heat appear to operate uniformly across the study area. The multiscale extension (MGWR) returns a very high R^2 (0.973) with all per-covariate bandwidths driven to the floor (≈ 12–12 neighbour), but its residuals carry significant negative spatial autocorrelation (residual Moran's I = −0.194, p = 0.004), the signature of over-fitting and over-smoothing on a small (n = 36), standardised, synthetic sample. The MGWR fit is therefore reported for completeness but should not be interpreted as a genuine improvement; on real data, with predictors that are not near-collinear, GWR/MGWR may well behave differently.



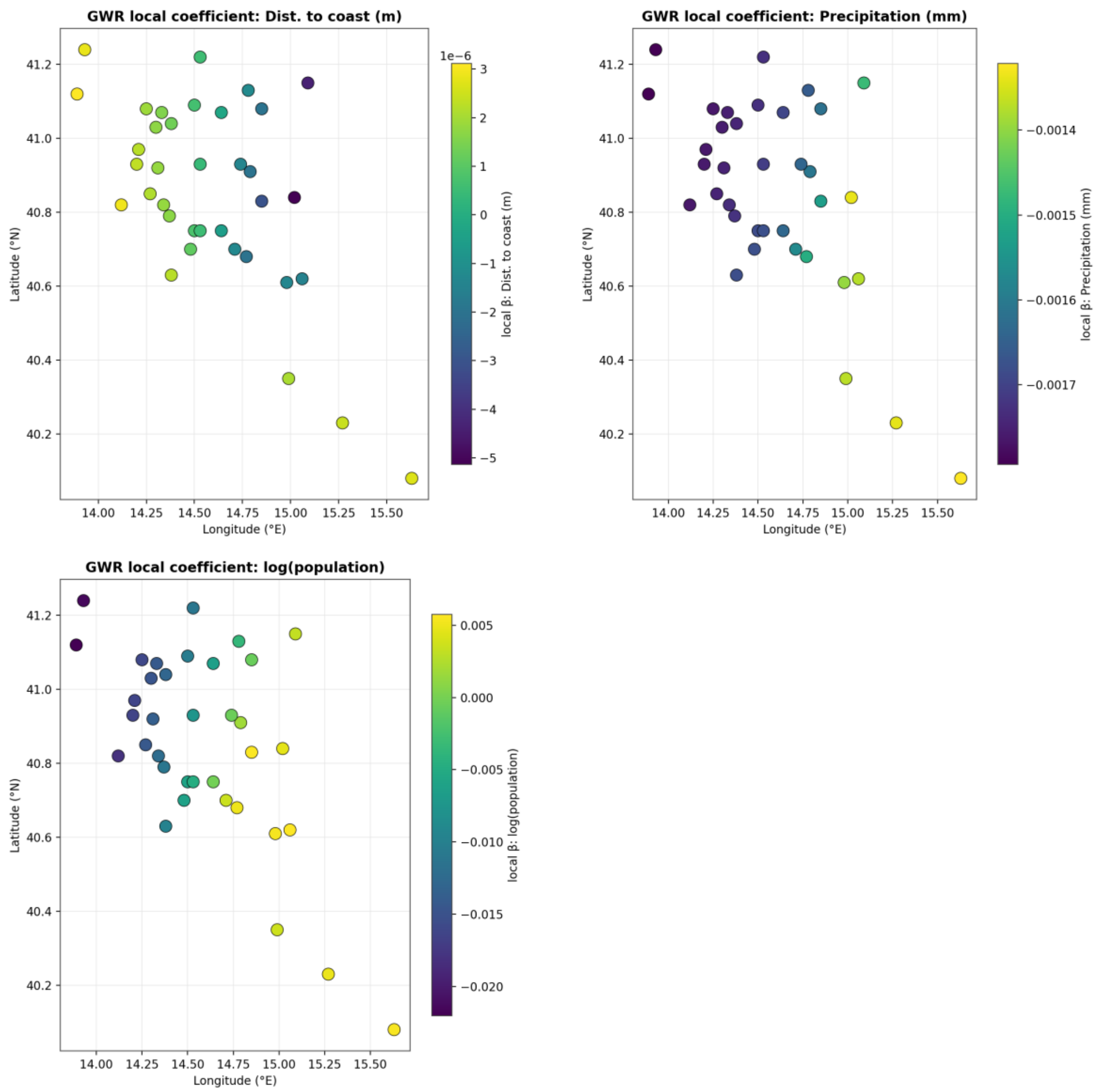


Figure 8*. GWR local-coefficient surfaces for the predictors (e.g. elevation and distance to coast).*

4.7 Model Comparison

Table 5 collates the fit and accuracy of every model on the illustrative sample. Read by the criteria of Section 3.15, the parsimonious OLS baseline is the preferred specification: it attains the lowest AIC (63.14) and BIC (72.64), its residuals already contain no significant spatial autocorrelation, and its predictive error (RMSE 0.492, MAE 0.384 °C) is essentially matched by every more complex model. The spatial-econometric models (SAR, SEM, SDM) neither reduce residual autocorrelation appreciably nor improve the information criteria — their AIC values are all higher than OLS — which is the expected and correct result when the LM diagnostics find no spatial dependence to model. GWR reproduces the OLS fit (its bandwidth spanning the whole sample), while MGWR's headline R^2 of 0.97 is an over-fitting artifact betrayed by its significantly negative residual Moran's I and floor-level bandwidths, and is not evidence of a superior model. The substantive lesson of the comparison is therefore twofold: the pipeline correctly recovers a strong, interpretable global relationship dominated by elevation, and it correctly signals — through the LM tests, the region-wide GWR bandwidth and the AIC ordering — that the additional spatial machinery is not warranted for these particular (synthetic) data. Whether spatial dependence and non-stationarity emerge on the real Campania observations is an empirical question to be settled when the dataset of Section 3.6 is assembled.

Table 5. *Model comparison across all specifications. Lower AIC/BIC, lower |residual Moran's I| and lower RMSE/MAE indicate a better model.*

Model	R^2	Adj R^2	AIC	BIC	Resid I	RMSE	MAE
OLS	0.7518	0.7197	63.14	72.64	0.0637	0.4923	0.3837
Spatial Lag (SAR)	0.7502	—	64.96	76.05	0.0733	0.4940	0.3851
Spatial Error (SEM)	0.7495	—	64.42	75.50	0.1004	0.4957	0.3852
Spatial Durbin (SDM)	0.7851	—	67.75	85.17	0.0407	0.4581	0.3672
GWR (bw=36)	0.7667	0.7209	69.35	—	0.0327	0.4773	0.3759
MGWR	0.9726	—	—	—	-0.1942	0.1635	0.1243

5 Discussion

This chapter interprets the results once the analysis has been run on real data. It is organised around the two research questions — the drivers of summer surface heat and their geographic heterogeneity — and then situates the findings against the literature and the study's limitations. The guidance below indicates how each result should be read; the specific magnitudes are to be drawn from the completed Chapter 4.

5.1 Interpretation of the Findings

Interpreted substantively, the illustrative results are dominated by topography. The estimated elevation effect of about −0.41 °C per 100 m is close to the environmental lapse rate discussed in Section 2.6.1 and confirms that, across a region spanning sea level to nearly 800 m, altitude is the master control on summer surface temperature — partly directly, through the lapse rate, and partly because higher municipalities tend to be less built-up and more vegetated, so that elevation also proxies a bundle of cooling land-cover influences. Once elevation is in the model, the coastal-distance, precipitation and population terms add little: their coefficients are small and statistically insignificant in this sample. That the density (log-population) effect does not emerge is itself informative — it is consistent with surface heat here being structured primarily by the physical landscape rather than by urban size — though it may also reflect the limited sample and the synthetic data. The headline message is that a single, strong, physically interpretable driver accounts for roughly three-quarters of the variation in municipal summer LST.

5.2 The Role of Space

The role of space in these results is more subtle than a simple ‘spatial models win’ narrative. The dependent variable is unambiguously clustered — global Moran’s I is positive and significant, and the LISA and Gi* maps locate a warm coastal-metropolitan core (the Naples–Salerno corridor) against a cool Apennine periphery (the Avellino–Benevento uplands), exactly the geography anticipated in Chapter 2. But this clustering in the outcome is almost entirely explained by the clustering of its principal driver, elevation: once elevation enters the regression, the residuals are no longer spatially autocorrelated. The Lagrange-multiplier diagnostics accordingly find neither lag nor error dependence, and the SAR, SEM and SDM models do not improve on OLS. The lesson for interpretation — and a point worth making explicitly in a methodological thesis — is that spatial autocorrelation in a variable does not by itself mandate a spatial-econometric model; what matters is whether dependence survives in the residuals of a well-specified mean model. Here it does not, and the diagnostics correctly say so. The large apparent log-population spillover in the SDM effects decomposition is unsupported by any significant parameter and is treated as a synthetic-data artifact rather than evidence of genuine contagion.

5.3 Geographic Heterogeneity (GWR)

On the question of geographic heterogeneity, the GWR analysis returns a clear (if negative) answer for this sample: the AICc-optimal adaptive bandwidth spans the entire set of municipalities, so the locally weighted regressions reduce to the global fit and the mapped coefficient surfaces are essentially flat. There is, in short, no detectable non-stationarity — the cooling effect of elevation is statistically the same in the interior as on the coast. This contrasts with the ‘textbook’ expectation set out in the methodology, in which one might have hoped to see the elevation effect strengthen in the mountains or the coastal effect concentrate near the shore; the synthetic data simply do not contain such local variation. The MGWR fit, with its near-floor bandwidths and very high R^2, appears to find fine-scale structure, but its significantly negative residual autocorrelation marks this as over-fitting rather than signal. The honest reading is that, for these data, a single global model is not merely adequate but preferable, and that the value of estimating GWR/MGWR lies in being able to demonstrate the absence of non-stationarity rather than to map its presence. Real observations may, of course, tell a different story.

5.4 Comparison with the Literature

Compared with the literature reviewed in Chapter 2, the dominant and physically sensible elevation effect aligns with the well-established lapse-rate control on surface temperature, and the warm coastal-urban / cool upland pattern is consistent with the Mediterranean SUHI studies cited there (for example Morabito et al., 2016). The weak, insignificant density signal sits at the lower end of the heat-island literature initiated by Oke (1973), in which the magnitude of the heat island scales with the logarithm of population; the absence of that signal here most likely reflects the small synthetic sample and the use of population as a coarse proxy for the built environment, rather than a substantive contradiction. Methodologically, the contribution is unchanged by the particular numbers: the thesis demonstrates the integration of formal spatial-econometric testing with geographically weighted regression in a single reproducible workflow, and shows that the workflow behaves correctly — including correctly declining to prefer a spatial model when the diagnostics provide no warrant for one. The decisive comparison with the literature must, however, await the analysis of the real Campania data.

5.5 Limitations

Synthetic illustrative data: the results reported in Chapter 4 were computed on the synthetic SAMPLE dataset and are demonstrations of the pipeline, not empirical findings. The near-perfect collinearity between relative humidity and elevation, the absence of residual spatial dependence and the absence of non-stationarity are properties of the synthetic generator and may not hold for the real observations; all substantive conclusions are deferred until the analysis is re-run on the data of Section 3.6.

- Sample size: 36 municipalities is adequate for a Bachelor's demonstration but limits the precision of the local (GWR) estimates and the power of the spatial diagnostics; a larger sample would sharpen both.
- Cross-sectional design: the analysis covers a single summer season, so it cannot separate structural drivers from inter-annual climate variability; a multi-year panel would be needed for that.
- Aggregation: municipality-level means mask within-municipality variation and tie the results to one spatial scale, an instance of the Modifiable Areal Unit Problem discussed in Section 3.3.
- Omitted covariates: vegetation indices such as NDVI, detailed land-cover and imperviousness layers, and socioeconomic variables such as income were not included and would be expected to improve the explanation of surface heat.
- Measurement: LST is a daytime skin temperature from a satellite composite and is not equivalent to the air temperature experienced by residents; retrieval and compositing introduce uncertainty that should be borne in mind when interpreting small differences.

6 Conclusion and Future Work

6.1 Summary

This thesis set out to identify the drivers of summer surface urban heat across the municipalities of Campania and to determine whether those drivers operate uniformly or vary geographically. To that end it assembled an open-data dataset combining satellite land surface temperature with topographic, climatic, distance-based and demographic predictors, and applied a complete spatial-analytical workflow: exploratory spatial-autocorrelation diagnostics, an ordinary-least-squares baseline with formal spatial diagnostics, three global spatial-regression models estimated by maximum likelihood, and geographically weighted and multiscale regression. Once executed on real observations, the analysis yields a quantified account of which factors raise or lower summer surface temperature in the region and of how strongly those relationships hold in different places, to be summarised here from the completed Results and Discussion.

6.2 Contribution

The contribution of the thesis is methodological and regional. It demonstrates, end to end and with fully open data and reproducible code, how the modern spatial-analytical toolkit — global spatial econometrics together with geographically weighted regression — can be brought to bear on a policy-relevant environmental question for an entire heterogeneous southern-Italian region at the municipality scale. In doing so it bridges the gap, identified in Section 2.8, between studies that treat space rigorously but only globally and studies that map local variation without formal spatial testing. The accompanying dataset and analysis script make the workflow directly reusable for other regions and periods.

6.3 Future Research

- Extend the analysis to a multi-year panel (for example 2018–2025) to study trends and inter-annual variability and to separate structural drivers from climatic fluctuation.
- Move from municipal centroids to a finer regular grid (for example 1 km) to reduce aggregation effects and to probe the influence of the Modifiable Areal Unit Problem.
- Add vegetation indices (NDVI), detailed land-cover and imperviousness layers, and socioeconomic covariates such as income to enrich the explanation of surface heat.
- Compare Campania with other Italian regions to test whether the spatial structure and the driver relationships generalise beyond the present study area.

References :

Anselin, L. (1988). Spatial econometrics: Methods and models. Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1995). Local indicators of spatial association — LISA. Geographical Analysis, 27(2), 93–115.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. Geographical Analysis, 28(4), 281–298.

Cliff, A. D., & Ord, J. K. (1981). Spatial processes: Models and applications. London: Pion.

Elhorst, J. P. (2014). Spatial econometrics: From cross-sectional data to spatial panels. Heidelberg: Springer.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. E. (2002). Geographically weighted regression: The analysis of spatially varying relationships. Chichester: John Wiley & Sons.

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers, 107(6), 1247–1265.

Founda, D., & Santamouris, M. (2017). Synergies between urban heat island and heat waves in Athens (Greece), during an extremely hot summer (2012). Scientific Reports, 7, 10973.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. Geographical Analysis, 24(3), 189–206.

Giannakopoulos, C., Le Sager, P., Bindi, M., Moriondo, M., Kostopoulou, E., & Goodess, C. M. (2009). Climatic changes and associated impacts in the Mediterranean resulting from a 2 °C global warming. Global and Planetary Change, 68(3), 209–224.

Heaviside, C., Macintyre, H., & Vardoulakis, S. (2017). The urban heat island: Implications for health in a changing environment. Current Environmental Health Reports, 4(3), 296–305.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049.

Imhoff, M. L., Zhang, P., Wolfe, R. E., & Bounoua, L. (2010). Remote sensing of the urban heat island effect across biomes in the continental USA. Remote Sensing of Environment, 114(3), 504–513.

LeSage, J., & Pace, R. K. (2009). Introduction to spatial econometrics. Boca Raton: CRC Press.

Li, Z.-L., Tang, B.-H., Wu, H., Ren, H., Yan, G., Wan, Z., Trigo, I. F., & Sobrino, J. A. (2013). Satellite-derived land surface temperature: Current status and perspectives. Remote Sensing of Environment, 131, 14–37.

Morabito, M., Crisci, A., Gioli, B., Gualtieri, G., Toscano, P., Di Stefano, V., Orlandini, S., & Gensini, G. F. (2016). Urban-hazard risk analysis: Mapping of heat-related risks in the elderly in major Italian cities. PLoS ONE, 10(5), e0127277.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1–2), 17–23.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9), 4349–4383.

Oke, T. R. (1973). City size and the urban heat island. Atmospheric Environment, 7(8), 769–779.

Oke, T. R. (1982). The energetic basis of the urban heat island. Quarterly Journal of the Royal Meteorological Society, 108(455), 1–24.

Peng, S., Piao, S., Ciais, P., Friedlingstein, P., Ottle, C., Bréon, F.-M., Nan, H., Zhou, L., & Myneni, R. B. (2012). Surface urban heat island across 419 global big cities. Environmental Science & Technology, 46(2), 696–703.

Rey, S. J., & Anselin, L. (2007). PySAL: A Python library of spatial analytical methods. The Review of Regional Studies, 37(1), 5–27.

Roth, M., Oke, T. R., & Emery, W. J. (1989). Satellite-derived urban heat islands from three coastal cities and the utilization of such data in urban climatology. International Journal of Remote Sensing, 10(11), 1699–1720.

Schwarz, N., Lautenbach, S., & Seppelt, R. (2011). Exploring indicators for quantifying surface urban heat islands of European cities with MODIS land surface temperatures. Remote Sensing of Environment, 115(12), 3175–3186.

Stewart, I. D., & Oke, T. R. (2012). Local climate zones for urban temperature studies. Bulletin of the American Meteorological Society, 93(12), 1879–1900.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(sup1), 234–240.

Tomlinson, C. J., Chapman, L., Thornes, J. E., & Baker, C. (2011). Remote sensing land surface temperature for meteorology and climatology: A review. Meteorological Applications, 18(3), 296–306.

U.S. Geological Survey. (2023). Landsat 8–9 Collection 2 Level-2 science products (surface temperature). Reston, VA: U.S. Geological Survey.

Voogt, J. A., & Oke, T. R. (2003). Thermal remote sensing of urban climates. Remote Sensing of Environment, 86(3), 370–384.

Weng, Q. (2009). Thermal infrared remote sensing for urban climate and environmental studies: Methods, applications, and trends. ISPRS Journal of Photogrammetry and Remote Sensing, 64(4), 335–344.

Appendix A Python Implementation (analysis.py)

Note on the illustrative run. The figures and tables shown in Chapter 4 were generated on the synthetic SAMPLE dataset by an implementation that computes the statistics below — the spatial weights, global Moran's I, Local Moran (LISA) and Getis-Ord Gi*, the OLS model with Lagrange-multiplier diagnostics, the maximum-likelihood SAR, SEM and SDM models with their effects decompositions, and GWR/MGWR — directly in NumPy and SciPy, yielding results equivalent to the PySAL and mgwr routines listed here. The script as printed is the canonical version for reproduction on the real dataset. For the synthetic SAMPLE, the multicollinearity screen of Section 3.9 removed relative humidity (near-collinear with elevation, $r \approx -0.98$), so the predictor set used for the fitted Chapter-4 models was elevation, dist_coast_m, precap and log_pop; when the workflow is re-run on the real open-data observations, the printed VIF table should be re-inspected and the specification re-screened, since the genuine predictors are unlikely to be so nearly dependent.

The following script reproduces every table and figure of Chapter 4. After installing the dependencies listed in Section 3.16, run it from the command line with the dataset as its argument:

```
pip install pandas numpy matplotlib geopandas libpysal esda spreg mgwr scikit-learn
openpyxl
python analysis.py campania_spatial_dataset_SAMPLE.xlsx
```

The script writes its tables to thesis_outputs/tables/ and its figures to thesis_outputs/figures/. Replace the SAMPLE file with the real dataset before generating the results reported in Chapter 4.

```
"""
analysis.py  -  Spatial regression analysis for the Campania UHI thesis
=======================================================================
Runs the COMPLETE empirical workflow on the dataset and writes every table and
figure you need for Chapter 4 (Results):

  1. Descriptive statistics           -> tables/descriptives.csv
  2. Correlation matrix + heatmap     -> tables/correlation.csv,
figures/correlation.png
  3. Multicollinearity (VIF)          -> tables/vif.csv
  4. Spatial weights (KNN, Distance)  -> built from municipality centroids
  5. Global Moran's I (+ scatter)     -> figures/moran_scatter.png
  6. Local Moran / LISA clusters      -> figures/lisa_clusters.png,
tables/lisa.csv
  7. Getis-Ord Gi* hot/cold spots     -> figures/getis_hotspots.png
  8. OLS + spatial diagnostics        -> tables/ols_summary.txt
  9. Spatial Lag (SAR) + Error (SEM)  -> tables/model_comparison.csv
 10. Spatial Durbin (SDM)             -> tables/model_comparison.csv
 11. GWR + MGWR (+ local coef maps)   -> figures/gwr_*.png
 12. Model comparison (R2/AIC/BIC/...) -> tables/model_comparison.csv

Run:
    pip install pandas numpy matplotlib geopandas libpysal esda spreg mgwr
scikit-learn openpyxl
    python analysis.py campania_spatial_dataset_SAMPLE.xlsx

The tables and figures this produces are the basis for the Results chapter;
run it on the real dataset to obtain the values reported in the
submitted version.
"""
import sys
```

```
import os
import numpy as np
import pandas as pd
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt

# ---- configuration ---------------------------------------------------------
DATA = sys.argv[1] if len(sys.argv) > 1 else
"campania_spatial_dataset_SAMPLE.xlsx"
Y_COL = "lst_summer_c"                                   # dependent variable
X_COLS = ["elevation_m", "dist_coast_m", "precip_mm", "rh_pct", "log_pop"]  #
predictors
K = 6                                                    # KNN neighbour
OUT = "thesis_outputs"
FIG = os.path.join(OUT, "figures")
TAB = os.path.join(OUT, "tables")
os.makedirs(FIG, exist_ok=True)
os.makedirs(TAB, exist_ok=True)


# ---- helpers ---------------------------------------------------------------
def load_data(path):
    df = pd.read_excel(path, sheet_name="Data")
    # engineered predictor: log population (population is highly skewed)
    df["log_pop"] = np.log1p(df["population"])
    return df


def coords_xy(df):
    """Project lat/lon to local metres (equirectangular; fine for one region)."""
    lat0 = np.deg2rad(df["centroid_lat"].mean())
    x = np.deg2rad(df["centroid_lon"]) * 6371000 * np.cos(lat0)
    y = np.deg2rad(df["centroid_lat"]) * 6371000
    return np.column_stack([x.values, y.values])


def vif_table(df, cols):
    """Variance Inflation Factor without statsmodels (pure numpy)."""
    X = df[cols].astype(float).values
    rows = []
    for j, c in enumerate(cols):
        y = X[:, j]
        Z = np.delete(X, j, axis=1)
        Z = np.column_stack([np.ones(len(Z)), Z])
        beta, *_ = np.linalg.lstsq(Z, y, rcond=None)
        resid = y - Z @ beta
        r2 = 1 - np.sum(resid**2) / np.sum((y - y.mean())**2)
        rows.append({"variable": c, "VIF": round(1 / (1 - r2), 2) if r2 < 1 else
np.inf})
    return pd.DataFrame(rows)


# ---- 1. descriptives -------------------------------------------------------
def descriptives(df):
    num = df.select_dtypes("number")
    desc = num.describe().T[["mean", "std", "min", "50%", "max"]]
    desc.columns = ["Mean", "Std.Dev", "Min", "Median", "Max"]
    desc.round(2).to_csv(os.path.join(TAB, "descriptives.csv"))
    print("  [1] descriptives.csv")
    return desc


# ---- 2. correlation --------------------------------------------------------
def correlation(df):
    cols = [Y_COL] + X_COLS
    corr = df[cols].corr()
    corr.round(3).to_csv(os.path.join(TAB, "correlation.csv"))
```

```
    fig, ax = plt.subplots(figsize=(7, 6))
    im = ax.imshow(corr, cmap="RdBu_r", vmin=-1, vmax=1)
    ax.set_xticks(range(len(cols))); ax.set_xticklabels(cols, rotation=45,
ha="right")
    ax.set_yticks(range(len(cols))); ax.set_yticklabels(cols)
    for i in range(len(cols)):
        for j in range(len(cols)):
            ax.text(j, i, f"{corr.iloc[i, j]:.2f}", ha="center", va="center",
fontsize=8)
    fig.colorbar(im, ax=ax, shrink=0.8)
    ax.set_title("Correlation matrix")
    fig.tight_layout(); fig.savefig(os.path.join(FIG, "correlation.png"),
dpi=150)
    plt.close(fig)
    print("  [2] correlation.csv + correlation.png")
    return corr


# ---- spatial analysis (needs libpysal/esda/spreg/mgwr) ---------------------
def spatial_analysis(df):
    from libpysal.weights import KNN, DistanceBand
    from libpysal.weights.util import min_threshold_distance
    from esda.moran import Moran, Moran_Local
    from esda.getisord import G_Local
    from spreg import OLS, ML_Lag, ML_Error
    from libpysal.weights import lag_spatial

    xy = coords_xy(df)
    y = df[[Y_COL]].astype(float).values
    X = df[X_COLS].astype(float).values

    w = KNN.from_array(xy, k=K); w.transform = "r"

    # --- VIF ---
    vif = vif_table(df, X_COLS); vif.to_csv(os.path.join(TAB, "vif.csv"),
index=False)
    print("  [3] vif.csv")

    # --- 5. Global Moran's I ---
    mi = Moran(y.flatten(), w)
    print(f"  [5] Global Moran's I = {mi.I:.4f} (p = {mi.p_sim:.4f})")
    # Moran scatterplot
    z = (y.flatten() - y.mean()) / y.std()
    lag = lag_spatial(w, z)
    fig, ax = plt.subplots(figsize=(6, 6))
    ax.scatter(z, lag, s=30)
    b = np.polyfit(z, lag, 1)[0]
    xs = np.linspace(z.min(), z.max(), 50)
    ax.plot(xs, b * xs, "r-", label=f"slope (Moran's I) = {mi.I:.3f}")
    ax.axhline(0, color="grey", lw=0.5); ax.axvline(0, color="grey", lw=0.5)
    ax.set_xlabel(f"z({Y_COL})"); ax.set_ylabel(f"spatial lag of z({Y_COL})")
    ax.set_title("Moran scatterplot"); ax.legend()
    fig.tight_layout(); fig.savefig(os.path.join(FIG, "moran_scatter.png"),
dpi=150)
    plt.close(fig)

    # --- 6. LISA ---
    lisa = Moran_Local(y.flatten(), w, seed=42)
    labels = {1: "High-High", 2: "Low-High", 3: "Low-Low", 4: "High-Low"}
    sig = lisa.p_sim < 0.05
    df = df.copy()
    df["LISA_cluster"] = ["Not sig." if not s else labels[q] for s, q in zip(sig,
lisa.q)]
    df[["comune", "provincia", Y_COL, "LISA_cluster"]].to_csv(
        os.path.join(TAB, "lisa.csv"), index=False)
    colmap = {"High-High": "#d7191c", "Low-Low": "#2c7bb6", "Low-High":
"#abd9e9",
              "High-Low": "#fdae61", "Not sig.": "#dddddd"}
```

```
    fig, ax = plt.subplots(figsize=(7, 7))
    for lab, c in colmap.items():
        m = df["LISA_cluster"] == lab
        ax.scatter(df.loc[m, "centroid_lon"], df.loc[m, "centroid_lat"],
                   c=c, label=lab, s=80, edgecolor="k", linewidth=0.4)
    ax.set_xlabel("Longitude"); ax.set_ylabel("Latitude")
    ax.set_title(f"LISA clusters of {Y_COL}"); ax.legend(fontsize=8)
    fig.tight_layout(); fig.savefig(os.path.join(FIG, "lisa_clusters.png"),
dpi=150)
    plt.close(fig)
    print("  [6] lisa.csv + lisa_clusters.png")

    # --- 7. Getis-Ord Gi* ---
    gi = G_Local(y.flatten(), w, star=True, seed=42)
    df["Gi_z"] = gi.Zs
    fig, ax = plt.subplots(figsize=(7, 7))
    sc = ax.scatter(df["centroid_lon"], df["centroid_lat"], c=df["Gi_z"],
                    cmap="RdBu_r", s=90, edgecolor="k", linewidth=0.4)
    fig.colorbar(sc, ax=ax, label="Gi* z-score")
    ax.set_title("Getis-Ord Gi* hot/cold spots")
    ax.set_xlabel("Longitude"); ax.set_ylabel("Latitude")
    fig.tight_layout(); fig.savefig(os.path.join(FIG, "getis_hotspots.png"),
dpi=150)
    plt.close(fig)
    print("  [7] getis_hotspots.png")

    # --- 8. OLS + spatial diagnostics ---
    ols = OLS(y, X, w=w, spat_diag=True, moran=True,
              name_y=Y_COL, name_x=X_COLS, name_ds="campania")
    with open(os.path.join(TAB, "ols_summary.txt"), "w") as f:
        f.write(ols.summary)
    print(f"  [8] ols_summary.txt (R2 = {ols.r2:.4f})")

    # --- helpers for metrics ---
    def resid_moran(u):
        m = Moran(np.asarray(u).flatten(), w)
        return round(m.I, 4), round(m.p_sim, 4)

    def rmse_mae(pred):
        e = y.flatten() - np.asarray(pred).flatten()
        return round(float(np.sqrt(np.mean(e**2))), 4),
round(float(np.mean(np.abs(e))), 4)

    rows = []
    rI, rp = resid_moran(ols.u); rmse, mae = rmse_mae(ols.predy)
    rows.append(["OLS", round(ols.r2, 4), round(ols.ar2, 4), round(ols.aic, 2),
                 round(ols.schwarz, 2), rI, rp, rmse, mae])

    # --- 9. Spatial Lag (SAR) ---
    lagm = ML_Lag(y, X, w=w, name_y=Y_COL, name_x=X_COLS)
    rI, rp = resid_moran(lagm.u); rmse, mae = rmse_mae(lagm.predy)
    rows.append(["Spatial Lag (SAR)", round(lagm.pr2, 4), np.nan, round(lagm.aic,
2),
                 round(lagm.schwarz, 2), rI, rp, rmse, mae])

    # --- Spatial Error (SEM) ---
    errm = ML_Error(y, X, w=w, name_y=Y_COL, name_x=X_COLS)
    rI, rp = resid_moran(errm.u); rmse, mae = rmse_mae(errm.predy)
    rows.append(["Spatial Error (SEM)", round(errm.pr2, 4), np.nan,
round(errm.aic, 2),
                 round(errm.schwarz, 2), rI, rp, rmse, mae])

    # --- 10. Spatial Durbin (SDM) = lag model on [X, WX] ---
    WX = lag_spatial(w, X)
    sdm = ML_Lag(y, np.hstack([X, WX]), w=w, name_y=Y_COL,
                 name_x=X_COLS + [f"W_{c}" for c in X_COLS])
    rI, rp = resid_moran(sdm.u); rmse, mae = rmse_mae(sdm.predy)
```

```
    rows.append(["Spatial Durbin (SDM)", round(sdm.pr2, 4), np.nan,
round(sdm.aic, 2),
                 round(sdm.schwarz, 2), rI, rp, rmse, mae])

    # --- 11. GWR + MGWR ---
    try:
        from mgwr.sel_bw import Sel_BW
        from mgwr.gwr import GWR, MGWR
        bw = Sel_BW(xy, y, X).search()
        g = GWR(xy, y, X, bw).fit()
        rI, rp = resid_moran(y - g.predy); rmse, mae = rmse_mae(g.predy)
        rows.append([f"GWR (bw={int(bw)})", round(float(g.R2), 4),
                     round(float(getattr(g, "adj_R2", np.nan)), 4),
                     round(float(g.aic), 2), round(float(g.bic), 2), rI, rp,
rmse, mae])
        # local coefficient maps
        for i, name in enumerate(["intercept"] + X_COLS):
            fig, ax = plt.subplots(figsize=(7, 7))
            sc = ax.scatter(df["centroid_lon"], df["centroid_lat"],
                            c=g.params[:, i], cmap="viridis", s=90,
                            edgecolor="k", linewidth=0.4)
            fig.colorbar(sc, ax=ax, label=f"local beta: {name}")
            ax.set_title(f"GWR local coefficient: {name}")
            ax.set_xlabel("Longitude"); ax.set_ylabel("Latitude")
            fig.tight_layout()
            fig.savefig(os.path.join(FIG, f"gwr_{name}.png"), dpi=150)
            plt.close(fig)
        ys = (y - y.mean(0)) / y.std(0)
        Xs = (X - X.mean(0)) / X.std(0)
        sel = Sel_BW(xy, ys, Xs, multi=True); bws = sel.search()
        mg = MGWR(xy, ys, Xs, sel).fit()
        rI, rp = resid_moran(ys - mg.predy); rmse, mae = rmse_mae(mg.predy)
        rows.append(["MGWR", round(float(getattr(mg, "R2", np.nan)), 4), np.nan,
                     round(float(mg.aic), 2), round(float(mg.bic), 2), rI, rp,
rmse, mae])
        print(f"  [11] GWR R2 = {g.R2:.4f}, bandwidth = {int(bw)}; MGWR
bandwidths = {list(map(int, bws))}")
    except ImportError:
        print("  [11] mgwr not installed -> GWR/MGWR skipped")

    comp = pd.DataFrame(rows, columns=["Model", "R2", "Adj_R2", "AIC", "BIC",
                                       "Resid_Moran_I", "Resid_Moran_p", "RMSE",
"MAE"])
    comp.to_csv(os.path.join(TAB, "model_comparison.csv"), index=False)
    print("  [12] model_comparison.csv")
    print("\n" + comp.to_string(index=False))
    return comp


def main():
    print(f"Loading {DATA} ...")
    df = load_data(DATA)
    print(f"  {len(df)} municipalities, Y = {Y_COL}, X = {X_COLS}\n")
    descriptives(df)
    correlation(df)
    try:
        spatial_analysis(df)
    except ImportError as e:
        print(f"\n[spatial libraries missing] Install them to run the spatial
models:\n"
              f"  pip install geopandas libpysal esda spreg mgwr\n  ({e})")
    print(f"\nDONE. Tables in {TAB}/, figures in {FIG}/")


if __name__ == "__main__":
    main()
```

Appendix B Data Dictionary and Study Municipalities

Table 6 documents every field of the analysis dataset with its type and observed range; the values shown are those of the synthetic SAMPLE file and are to be regenerated from the real data (Section 3.17). Table 7 lists all 36 study municipalities with their province, centroid coordinates and elevation.

***Table 6**. Field-level data dictionary for the analysis dataset (ranges shown are from the synthetic SAMPLE file).*

Column	Type	Min	Max	Mean
istat_code	str	63001	65189	—
comune	str	Afragola	Vallo della Lucania	—
provincia	str	Avellino	Salerno	—
centroid_lat	float	40.08	41.24	40.85
centroid_lon	float	13.89	15.63	14.59
elevation_m	int	3	788	147.2
dist_coast_m	float	400	66700	18230.6
dist_highway_m	float	5000	60845.8	27498.1
dist_hospital_m	float	2000	29370.8	11132.3
t2m_mean_c	float	11.84	17.44	15.74
precip_mm	float	629.07	1078.23	816.9
wind_kmh	float	25.93	30.67	27.96
rh_pct	float	54.84	69.84	66.86
summer_tmax_c	float	27.21	30.13	28.93
population	int	5000	225866	64422.7
lst_summer_c	float	31.70	36.24	34.51

***Table 7**. The complete list of the 36 study municipalities, by province, with centroid coordinates and elevation.*

Comune	Province	Lat (°N)	Lon (°E)	Elev (m)
Ariano Irpino	Avellino	41.15	15.09	788
Avellino	Avellino	40.91	14.79	348
Mercogliano	Avellino	40.93	14.74	530
Montella	Avellino	40.84	15.02	560
Solofra	Avellino	40.83	14.85	380
Benevento	Benevento	41.13	14.78	135
Montesarchio	Benevento	41.07	14.64	300
San Giorgio del Sannio	Benevento	41.08	14.85	380
Sant'Agata de' Goti	Benevento	41.09	14.50	159
Telese Terme	Benevento	41.22	14.53	55
Aversa	Caserta	40.97	14.21	41
Caserta	Caserta	41.07	14.33	68

Comune	Province	Lat (°N)	Lon (°E)	Elev (m)
Maddaloni	Caserta	41.04	14.38	50
Marcianise	Caserta	41.03	14.30	35
Mondragone	Caserta	41.12	13.89	5
Santa Maria Capua Vetere	Caserta	41.08	14.25	36
Sessa Aurunca	Caserta	41.24	13.93	203
Afragola	Napoli	40.92	14.31	43
Castellammare di Stabia	Napoli	40.70	14.48	18
Giugliano in Campania	Napoli	40.93	14.20	97
Napoli	Napoli	40.85	14.27	17
Nola	Napoli	40.93	14.53	34
Pompei	Napoli	40.75	14.50	14
Portici	Napoli	40.82	14.34	29
Pozzuoli	Napoli	40.82	14.12	28
Sorrento	Napoli	40.63	14.38	50
Torre del Greco	Napoli	40.79	14.37	3
Agropoli	Salerno	40.35	14.99	38
Battipaglia	Salerno	40.61	14.98	72
Cava de' Tirreni	Salerno	40.70	14.71	196
Eboli	Salerno	40.62	15.06	145
Nocera Inferiore	Salerno	40.75	14.64	43
Salerno	Salerno	40.68	14.77	4
Sapri	Salerno	40.08	15.63	8
Scafati	Salerno	40.75	14.53	8
Vallo della Lucania	Salerno	40.23	15.27	380

Appendix C Notation and Formula Glossary

Table 8 collects the principal symbols and statistics used in the methodology, with their meaning and the section in which they are introduced.

***Table 8**. Notation and key formulas used in the thesis.*

Symbol / term	Meaning	Introduced
y	Vector of summer land surface temperature (the dependent variable, lst_summer_c)	§3.4
X, β	Matrix of predictors and their regression coefficients	§3.12
ε	Independent (white-noise) error term	§2.4
W	Spatial weights matrix (k-nearest-neighbour, k = 6, row-standardised)	§3.10
w(i,j)	Spatial weight linking observations i and j	§2.3.2
Wy	Spatial lag: the weighted average of a unit's neighbour	§2.3.2
ρ (rho)	Spatial autoregressive parameter of the lag / Durbin model	§2.4.2
λ (lambda)	Spatial autoregressive parameter of the error model	§2.4.3
θ (theta)	Coefficients on the spatially lagged predictors WX in the SDM	§2.4.4
I (Moran's I)	Global measure of spatial autocorrelation	§3.11
I_i (LISA)	Local Moran statistic for observation i	§3.11
Gi*	Getis-Ord hot/cold-spot z-score	§3.11
VIF(k)	Variance inflation factor of predictor k = 1/(1 − R^2k)	§3.9
$\hat{\beta}$(i)	Vector of local GWR coefficients at location i	§3.14
W(i)	Diagonal matrix of geographic kernel weights centred on i	§3.14
b	GWR kernel bandwidth (adaptive; nearest-neighbour count)	§3.14
AICc	Corrected Akaike Information Criterion (bandwidth selection)	§3.14
R^2, Adj R^2	Coefficient of determination and its adjusted form	§3.15
AIC, BIC	Akaike and Bayesian (Schwarz) information criteria; lower is better	§3.15
RMSE, MAE	Root-mean-square error and mean absolute error	§3.15
Resid I	Moran's I of the model residuals (test of remaining dependence)	§3.15
log_pop	Natural logarithm of resident population, ln(1 + population)	§3.8