Grup d'Estadística espacial i temporal
en Epidemiologia i Medi Ambient

Geostatistical computing of acoustic maps in the presence of barriers

Antonio López-Quílez, Antonio.Lopez@uv.es
Facundo Muñoz, Facundo.Munoz@uv.es

Technical Report: **GEeitEma.GuenMap .#5-2008**

Burjassot, september 20, 2008.

Geostatistical computing of acoustic maps in the presence of barriers[*]

Antonio López-Quílez Facundo Muñoz

Abstract

Acoustic maps are the main diagnostic tools used by authorities for addressing the growing problem of urban acoustic contamination. Geostatistics models phenomena with spatial variation, but restricted to homogeneous prediction regions. The presence of barriers such as buildings introduces discontinuities in prediction areas. In this paper we investigate how to incorporate information of a geographical nature into the process of geostatistical prediction. In addition, we study the use of a Cost-Based distance to quantify the correlation between locations.

keywords: non-Euclidean geostatistics, computational methods, acoustic maps, Cost-Based distance ,GIS

1 Introduction

Acoustic contamination in urban areas is becoming increasingly considered as a public health topic by authorities [1]. Noise maps are the diagnostic tools used for planning prevention and correctional measures. Noise maps represent, for each location, the mean noise level over a timespan, expressed on an appropriate scale.

Making a single measurement of noise level is not trivial. There are many restrictions to take into account, and it is an operation that takes no less than $15 - 30$ minutes of a qualified operator's work. In short, observations are expensive.

This led to another approach based on the simulation of deterministic models of noise diffusion which make use of a digitalized model of the city together with a number of traffic parameters for each road. But uncertainty and variability of parameters and simplifications in the model propagates error over thousands of iterations in an unknown and uncontrolled way.

Geostatistics provides a set of statistical tools specifically designed for spatial problems, in which prediction is required over a

[*] Supported in part by Ministerio de Ciencia e Innovación, through Grant CIT-310200-2007-85

region of interest where some observations have been taken. Predictions are based on an underlying statistical model that can take additional information into account as explanatory variables. In addition, the prediction error can be estimated based on the propagation of uncertainty.

The main drawback with geostatistics is that it assumes the area of interest to be a homogeneous, unrestricted region. However, it is clear that buildings and urban infrastructure represent restrictions or barriers to noise flow.

In this study we develop a methodology for overcoming this problem, taking advantage of modern Geographical Information Systems (GIS). We propose the use of a Cost-Based distance to quantify the correlation between locations. In this way we take into account the heterogeneous configuration of the environment.

We used `GRASS GIS` [2] for geographical analysis, and `R` [3] for geostatistical computation. Both are open source, free, powerful, flexible and customizable software. In addition, they communicate with each other easily through a library called `spgrass6` [4], and they provide scripting capabilities so automatization is possible. The `geoR` [5] package implements most geostatistical methods in `R`. We adapted some of its algorithms to implementing geostatistical models with non-Euclidean distances.

In Section 2 we present and define the so-called Cost-Based distance and compare it with the Euclidean distance. We show that this is a generalization that overcomes the classical "homogeneous and unrestricted region" constraint. We also outline the algorithm we developed to automatically compute this type of distance.

In Section 3 we explain the processing of geographical information, emphasizing the use of Cost-Based distances to relevant objects as explanatory variables.

In Section 4 we provide a brief review of classical geostatistical theory, and explain the modifications needed for implementing geostatistical analysis with Cost-Based distance.

In Section 5 we briefly outline the whole process and in Section 6 we present an example of its application to the problem of noise mapping. We make final comments and conclusions in Section 7.

2 Cost-Based Distance

2.1 Motivation and concept

Methods for spatial data analysis have typically been applied to convex subsets of $\mathbb{R}^2$ [6]. In this situation it is sensible to think of

Euclidean distance as the natural argument for a correlation function. However, the presence of barriers within the region of interest changes things. Imagine two locations at a given (Euclidean) distance such that they are significatively correlated, because of underlying relevant factors affecting both of them. Now put a barrier between them that blocks or absorbs the effect of the underlying factors. This obviously pulls the correlation down. So when barriers exist, the correlation depends on something other than the Euclidean distance, which therefore cannot account for correlation by itself.

A natural extension is to associate the correlation between two locations with the minimum distance that has to be traveled without crossing any barriers (see Fig. 1). Note that when there is no barrier at all, this reduces to a Euclidean distance. Little, Edwards and Porter, who worked with contaminants in estuaries, illustrated this by saying that distance could be measured *as the crow flies, or as the fish swims* [7].

There are more general situations where barriers are not absolute, but regions are harder (or easier) to cross. For example, a fungus in a field will easily spread over fertile, warm and protected portions of land. In contrast, it will spread with more difficulty over exposed and rocky areas. This heterogeneity can be modeled with a *Cost surface* representing how hard it is to cross a given portion of area. And accordingly, the correlation between two locations should be associated with the minimum-cost path connecting them. Formally:

Definition 2.1 (Cost Surface)**.** A function over the region of interest with values in the non-negative real numbers, such that the value at a given location is interpreted as the *cost density* at that point.

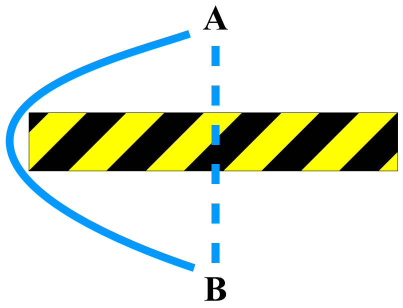


Figure 1: Cost-Based distance (continuous) vs. Euclidean distance (dotted).

This is the tool we use to represent every relevant factor affecting correlation. In particular, in this surface are synthesized the Euclidean distance and the environmental configuration. The Cost surface is not necessarily continuous, nor even bounded. For example, barriers are regions with infinite cost. It should be theoretically defined over all the plane, since the optimum path must be found among all possible paths. However, in practice it is enough to define it over a region covering all relevant locations, by arguing that all minimum-cost paths must lie within the region.

Any path connecting two locations has an associated cost:

Definition 2.2 (Cost of a path)**.** Given a Cost surface, every path lying within the working region has an associated cost that is computed by integrating the Cost surface along it.

Definition 2.3 (Cost-Based distance)**.** Given a Cost surface, the Cost-Based distance between two locations is defined as the cost of the minimum-cost path connecting them.

In this framework, the standard geostatistical assumptions where the region is homogeneous is a particular case where the Cost surface is a constant 1-valued surface, and therefore the minimum-cost path between two given locations is the *straight line* connecting them, hence the Cost-Based distance equals the Euclidean distance. Also, the more general situation with barriers in the working region is another particular case where the Cost surface takes the value 1 over non-barrier areas and the value ∞ over barrier areas, therefore the Cost-Based distance equals the minimum distance needing to be traveled without crossing any barriers, as was required.

2.2 Relationship with Euclidean distance

A natural question to ask is whether using Cost-Based distances makes a big difference or not. This will depend on how *different* (in some sense) the Cost surface is from a constant surface, which in turn depends on the *geometry* of the environment.

In an urban environment, buildings and infrastructure may well act as barriers for many response variables. Examples are pollutants, noise, light or anything, in general, that flows through air and is blocked by walls. Often, Euclidean and Cost-Based distances will not differ too much. For example, for every pair of points along the same road, both distances will coincide. However, it is not hard to find situations where the two types of distance are very different.

Example 2.4. In Fig. 2, consider the distance (both types) between the red dot and four points labeled A, B, C and D. It can be seen by

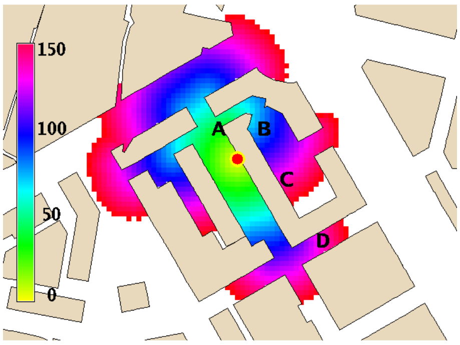


Figure 2: Partial distance map (in meters) associated to the red dot.

the naked eye that A and B are at the same Euclidean distance from the red dot; however, B is twice as far as A in Cost-Based distance terms. Conversely, D is twice as far as C, while they are at the same Cost-Based distance from the red dot.

2.3 Computation

In geostatistics, two distance matrices are implicitly used. One holds the distances between observation points. This is a symmetric square matrix, since the Euclidean distance from A to B is the same as the converse. The second matrix holds the distances between observation points and the prediction location(s), so it is an $n(\text{observations}) \times m(\text{locations})$ sized matrix.

Computing both matrices reduces to the general problem of calculating the distance matrix between two sets of points (which we will call the *from* and *to* sets). For the first matrix, the *from* and *to* sets are both the set of observations, while for the second they are the observation and the prediction location sets respectively.

Our solution to this problem is based on the computation of *distance maps*.

Definition 2.5 (Distance map for a given location)**.** A map that, for every point, represents its Cost-Based distance to the given location.

The color map in Fig. 2 is a partial representation of the distance map associated with the red dot. In this way, we can know the Cost-Based distance to the red dot for points A, B, C and D just by looking at the value of the distance map at their respective locations.

Ingredients for computation are:

- **A Coordinate Reference System (CRS), and a working region.** Every entity in our model must be (geo)referenced in the same system so we can measure relative distances. Besides, we need a finite working region of interest where observations and prediction locations are confined.

- **Coordinates of point(s) in the *from* and *to* sets.**
- **Cost surface.** A raster map (at some suitable resolution) covering the working region.

Resolution is a parameter of the computing process that affects both its speed and accuracy, in opposite directions. Theoretically, we use a continuous Cost surface, but in practice discretization is needed and this introduces error. The higher the resolution, the lower the error. On the other hand, it takes more time to compute the distance maps and they take up more disk space. Since we need as many maps as observation points, the difference in time and space can be very high. So resolution is a balance parameter which depends in particular on n, the number of observations.

The value assigned to a raster cell in the discretized Cost surface represents the integration of the continuous ideal surface over a path traversing that specific cell. Obviously the exact value depends on the relative position of the path with respect to the cell. Hence, the assigned value can only be an approximation, depending on the size of the cell, i.e. the resolution of the raster. As a result, the constant value of the discretized Cost surface is another parameter closely related to the resolution parameter. In our application we used the same value for both parameters. Resolution is 5 metres, so each raster cell represents a 5 m × 5 m portion of land, and the cost assigned to each (non-barrier) raster cell was also 5, meaning that traversing the cell requires a distance of approximately 5 metres long.

The distance maps are *cumulative cost surfaces*, computed by *expanding* from the associated point and accumulating the cost of each cell. There are several expansion strategies. For example, we can expand only to adjacent cells or we can expand also to diagonal cells, multiplying the cost of the cell by a correction factor. A more sophisticated alternative is also expanding as the Knight moves, which improves accuracy significantly. Over a constant Cost surface, the accumulated cost grows outwards in 4, 8 or 16 directions respectively (see Fig. 3).

Since we are working with georeferenced information, the natural environment to work within is a Geographical Information System(GIS). Most GIS software implements algorithms to compute

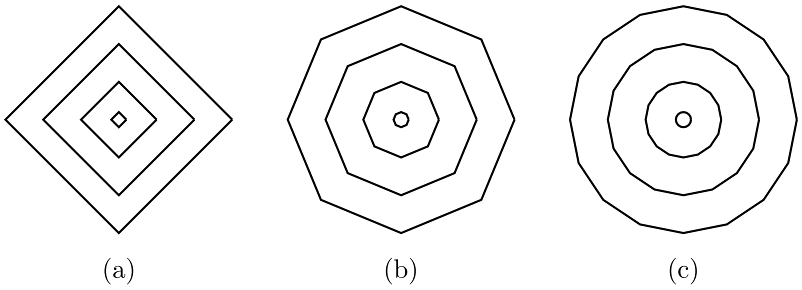


Figure 3: Level curves for a constant Cost surface with different expansion strategies. (a) Expanding to adjacent cells only (b) Expanding also in diagonals (c) Expanding also with Knight's move.

the cumulative cost surfaces, which is the most demanding step in terms of computational cost.

We implemented this process in GRASS GIS. The algorithm takes as input a raster map representing the Cost surface, a vector layer with the *from* set of points, and one or more vector layers with the *to* set(s). Null-valued cells in the cost raster are interpreted as infinite cost. It generates (temporal) raster distance maps (cumulative cost surfaces) for each of the *from* points, making use of the `r.cost` [8] GRASS base function. Then, for each of the *to* layers of points it generates as many columns as *from* points in the *to* layer's attribute table, filling them with values collected with corresponding values picked from the distance maps.

In summary:

```
FOR each point $A$ in the from set:
    – Compute its distance map $D$
    – FOR each point $B$ in the to set:
          Pick value of $D$ in position $B$
```

Using the observation layer as *from* and again the observation and prediction location layers as *to*, we get the two Cost-Based distance matrices required, with a single and automatic command.

3 GIS analysis

The first step in the GIS analysis stage is to create the prediction location layer, which will contain the points where predictions are required, and also the additional geographical information associated to each of the locations that will be used as covariates in the external trend estimation.

3.1 Prediction Locations

In the default case that no specific point is of particular interest, we assume that prediction is required over all possible areas within the working region. That is, we want to predict everywhere it makes sense. Deciding whether predicting somewhere makes sense or not requires geographical information that has to be provided in the form of a GIS layer(s).

A parameter of the process is the resolution of the prediction locations, i.e. how fine the grid is. This will affect the resolution of the final prediction maps. With this resolution we define a vectorial regular grid covering the region. We now make use of the input layer with either predictable or unpredictable areas (we should be able to compute one of them from available geographical information). In the former case we take the intersection with the predictable area, while in the latter case we subtract the unpredictable area from it. At this point, we have a tessellation of the predictable areas from where we can pick the centroids of the polygons as the prediction locations.

In our application, we had a layer with the buildings in the city, which were the places where we did not want to predict. So we started by creating a vector grid with the selected resolution and subtracted the building layer from it. The centroids of the resulting polygons were then picked as the prediction locations.

Fig. 4 shows the prediction area tessellation with centroids, once buildings (in beige) had been subtracted. It can be seen that this process generates a set of locations at the desired resolution, or higher in those cases closer to buildings.

Note that the operations carried out here are typical of GIS analysis: the creation of a vectorial grid, intersection/subtraction with another vectorial layer and centroid extraction. Since we are handling geographical information, GIS is the natural environment to work within at this point.

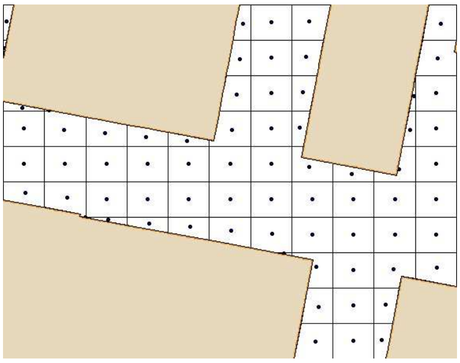


Figure 4: Automatic creation of prediction locations.

3.2 Incorporating additional geographical information

Geostatistical computations can be performed jointly with regression models, enabling the use of explanatory variables which may provide additional information about the response variable. In most environmental applications, for example, altitude turns out to be a very informative covariate.

Distances to relevant objects can also provide valuable information in some situations. For example, distances to sources of chemical waste disposals in contamination problems, or distances to noise sources in acoustics. Of course, for the reasons presented in previous sections, Cost-Based distances should be used in these models.

Some GIS analysis is required in order to make this possible. A distance map has to be computed for every relevant entity of interest. Finally, an iterative process, the analog to that described in Section 2, picks up the corresponding values for every observation and prediction location.

3.3 Representation of results

One final stage where GIS analysis is of particular use is, of course, the representation of the resulting maps. The outcomes of geostatistical techniques are prediction values and prediction error estimates for each of the prediction locations. These values are to be returned to the GIS as two new attributes of the prediction location layer. Recall that each prediction location originated as the centroid of a polygon, which was part of a tessellation of the predictable area. So it is sensible to assign those resulting values to each of the corresponding polygons, and to paint each polygon according to them, based on a common color scale. Again, both steps are easily performed on any GIS software.

4 Computing Cost-Based Geostatistics

4.1 Overview of geostatistical theory

Geostatistics is a branch of statistics that encompasses the techniques that apply to geographical analysis. It is said to have originated in the early 50's from the work of the South African mine engineer D. G. Krige [9], and developed and systematized by the work of Georges Matheron [10]. A classical reference on the field is [11].

There is a number of applications for geostatistical methods. The common underlying characteristic is that observations can be understood as a (partial) realization of a Stochastic Process over a continuous spatial region $D \subseteq \mathbb{R}^2$.

$$\{Z(\boldsymbol{s}) \;:\; \boldsymbol{s} \in D\}$$

This process is commonly assumed to be *Gaussian*, *isotropic* and *intrinsically stationary.* That is, for any collection of locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, with each $\boldsymbol{s}_i \in D$, the joint distribution of $\boldsymbol{Z} = \{Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)\}$ is multivariate normal, and the variance of the value differences between two locations depends only on the distance separating them. This variance is twice that known as the *semivariogram function* $\gamma(r)$.

$$\mathrm{Var}\left[Z(\boldsymbol{s}_1) - Z(\boldsymbol{s}_2)\right] = 2\gamma(r), \qquad r = \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$$

In order to define a legitimate model, the semivariogram function $\gamma(r)$ must be negative-definite. This condition imposes non-obvious constraints so as to ensure that, for any integer m, set of locations $\boldsymbol{s}_i$ and real constants λ_i, the linear combination $\sum_{i=1}^{m} \lambda_i Z(\boldsymbol{s}_i)$ will have non-negative variance. In practice, this is usually ensured by working within one of several standard classes of parametric models for $\gamma(r)$.

Estimation of the correlation structure is usually accomplished in terms of the semivariogram function by fitting the empirical semivariogram computed from observed data. There are a variety of methods for estimating the semivariogram function parameters. Our approach here is to use maximum likelihood methods, simultaneously fitting the mean function $\mu(\boldsymbol{s})$, possibly depending on additional covariates, and the parameters of the semivariogram function $\gamma(r)$.

Once a model is fitted to data, we are interested in prediction. There are many geostatistical approaches to this problem, but the most commonly used spatial prediction method is known as *kriging.*

Most methods use a weighted average of the sample values to generate the prediction; sample points near the prediction's location are given larger weights than those far away. Kriging determines these weights based on the semivariogram function.

Its popularity owes much to some nice properties of the kriging predictor. Being the *Best Linear Unbiased Predictor* (BLUP, in terms of quadratic error), it is remarkably robust to violations of model assumptions [12], and provides standard error predictions.

Kriging assumes that the observation vector $\boldsymbol{Z}$ is generated by an isotropic, intrinsically stationary Gaussian process with mean function $\mu(\boldsymbol{s}) = \beta_0 + \beta_1 f_1(\boldsymbol{s}) + \ldots + \beta_p f_p(\boldsymbol{s})$ and a known variogram function $\gamma(r)$, where the $f_i(\cdot)$ are functions of the spatial location $\boldsymbol{s}$ or explanatory variables associated to the locations.

The kriging predictor at a given site $\boldsymbol{s}_0$ is written as a linear combination of the data at the sampled sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$:

$$\hat{Z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i)$$

where $\lambda_1, \ldots, \lambda_n$ are chosen to minimize the mean squared prediction error

$$\mathrm{E}\big[\big(\hat{Z}(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)\big)^2\big]$$

subject to the unbiasedness constraint that $\mathrm{E}\big[\hat{Z}(\boldsymbol{s}_0)\big] = \mathrm{E}\big[Z(\boldsymbol{s}_0)\big]$.

This optimization problem leads to a constrained system of equations, with solution (see [11], or [13]):

$$\hat{Z}(\boldsymbol{s}_0) = \Big[\boldsymbol{\gamma} + \mathbf{X}(\mathbf{X}'\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}(\boldsymbol{x} - \mathbf{X}'\boldsymbol{\Gamma}^{-1}\gamma)\Big]'\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}$$

and prediction variance:

$$\sigma^2(\boldsymbol{s}_0) = \boldsymbol{\gamma}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - \big(\mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma} - 1\big)^2 / \big(\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}\big)$$

where $\boldsymbol{\gamma} = (\gamma(\|\boldsymbol{s}_1 - \boldsymbol{s}_0\|), \ldots, \gamma(\|\boldsymbol{s}_n - \boldsymbol{s}_0\|))'$, $\boldsymbol{x} = (f_0(\boldsymbol{s}_0), f_1(\boldsymbol{s}_0), \ldots, f_p(\boldsymbol{s}_0))'$, $\boldsymbol{\Gamma} = \big(\gamma(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)\big)$, and $\mathbf{X} = \big(f_{j-1}(\boldsymbol{s}_i)\big)$, being $f_0(\boldsymbol{s}) = 1 \quad \forall \boldsymbol{s}$.

4.2 Use of non-Euclidean distances

Geostatistics assumes that locations which are close together are more similar than locations that are far apart. The kriging predictor uses weights that are calculated according to the value of the variogram, which is a function of Euclidean distance. As was explained in Section 2, there are many situations where the argument

r of the variogram function is represented more naturally by the Cost-Based distance.

Various researchers have come to this conclusion since the work of Little et al. and Rathbun in the mid-90's in the field of geostatistical analysis in estuaries, where they found it natural to use "water distances". [14] and [15] followed them. Curriero showed that most traditional parametric covariance models are not valid for non-Euclidean distances. Hence, such distances cannot be used without proof of validity of the model. Other authors like [16], [17], [18], and recently [19] have explored different approaches such as moving window kernels or Multidimensional Scaling.

Geostatistical computations are better carried out within a powerful statistical environment, such as `R`. Instead of programming ad-hoc geostatistical algorithms, we adapted the `geoR` package by adding flexibility and enhancing it.

There are three major stages in classical geostatistical analysis computation that need to be adapted: empirical variogram computation, variogram model parameter fitting and the actual kriging prediction. Apart from observation data and prediction locations needed for standard kriging, we also need the two Cost-Based distance matrices previously computed, as explained in Section 2.

The empirical variogram is computed from the observation data only. It classifies pairs of observations into groups according to their distance, and then computes an estimator of the theoretical variogram value for that distance based on the differences between the observed values. In order to make a Cost-Based empirical variogram it is enough to make the initial classification based on the Cost-Based distance values given in the corresponding matrix, rather than calculating Euclidean distances. Note that this modification produces a different grouping of observation pairs. Therefore, variogram estimates will be different.

The variogram model parameter fitting is also made based on observation data only. It is typically accomplished through maximum likelihood methods, basically trying out many possible combinations iteratively and keeping the best. This implies computation of the covariance matrix for each combination being tested. All we need is to make sure that the covariance matrix is computed based on the Cost-Based distances provided by our previously computed matrix.

Finally, there is the kriging prediction. At this point, the covariance model is assumed to be known. But here again, we need to make sure that the covariance matrix of the observations is computed with the Cost-Based distances. In addition, the covariance between observation points and prediction locations are to be computed in order to make predictions. So this is when the second of the Cost-Based distance matrices is to be used.

5 Process Overview

One of the goals of the present study was to develop a computational tool to perform Cost-Based geostatistics with minimal user intervention. The whole process is outlined next for the particular case of absolute barriers.

GIS analysis

Inputs: geographical environment, barriers map, prediction resolution parameter.

1. Create a regular vector grid covering the whole working region with the specified resolution.
2. Crop the areas where barriers exists. The result is a tessellation of the prediction region.
3. Extract as points the centroids of each polygon from the tessellation.
4. Incorporating additional information as covariates: Cost-Based distance to relevant entities. Iterate over each one of the entities of interest.
   Inputs: entity and prediction region maps, observation and prediction location maps, cost computation resolution and maximum cost parameters.
   (a) Rasterization of the entity map with the given resolution.
   (b) Rasterization of the prediction region with the given resolution and with a raster value equal to the resolution size.
   (c) Computation of the distance map from the entity up to the maximum cost.
   (d) Pick up Cost-Based distances for both observation and prediction points.
5. Computing Cost-Based distance matrices: observations-observations and observations-locations
   Inputs: observation and location maps, Cost surface (barriers map).
   (a) Computation of the distance maps for each of the observation points.

(b) Pick up Cost-Based distances for both observation and prediction points and for each distance map.

Statistical prediction

Inputs: observations and prediction location maps, with attribute tables containing covariate values and Cost-distance matrices.

1. Selection of the regression model. Transformation and selection of covariates, interactions, etc.
2. Cost-Based empirical variogram computation.
3. Variogram model family selection and Cost-Based parameter estimation.
4. Cost-Based kriging prediction.
5. Return the prediction location map to the GIS with the attributes of prediction values and error estimates added.

Presentation of results

Inputs: prediction location maps with attribute tables containing prediction values and error estimates, tesselated prediction region, and everything else required for representation.

1. Transfer the attributes of prediction values and error estimates from the locations to the corresponding polygon of the tessellated prediction region map.
2. Configure thematic map options and show results.

6 Acoustic maps in the presence of barriers

6.1 Sample data

As a pilot application, we wanted to make a noise map of the Malilla neighborhood in the city of Valencia. We have made a set of 52 noise measurements in various points distributed over the neighborhood. A subset of measurements was taken very close to each other along two blocks: one of them situated on an avenue, the other on a small street. This was done to enable the estimation of the variogram in short-range distances.

The neighborhood is not homogeneous. In fact it has great contrasts, having big avenues very close to small and quiet streets. Traffic density explains much of the noise difference between locations, so every street in the neighborhood was classified according to its traffic density. Avenues are category 1 roads, medium traffic density roads are category 2, and quiet streets are category 3 roads. Fig. 5 shows the neighborhood with the observation locations and road classification.

6.2 Geostatistical setup

Buildings are considered non-transparent barriers for noise. Traffic noise spreads around the neighborhood between the buildings through open areas only. For this reason, we believe that Cost-Based distance to roads of higher traffic density are relevant explanatory variables. We consequently computed the Cost-Based distances to the closest road of each type in order to use them, once properly scaled, as covariates in a regression model. Figure 6 shows the distance maps (up to 100 m.) for each of the three road types.

We emphasize the use of Cost-Based distance instead of Euclidean distance for explanatory variables, since there are configurations where a point is very close to an avenue, though it is located behind a building that prevents the noise from reaching it. In addition, as explained in Section 2, Cost-Based distance explains the correlation between locations better than Euclidean distance. A natural question to ask is how different the values from the two types of distances are. Figure 7 shows the Cost-Based distance between all pairs of observations versus their Euclidean distance.

Note the step at approximately 95 metres. Pairs of observations under this distance threshold are those located along the same block, without any obstacle. Thus, there are no practical differences between the two types of distance. In contrast, from 95 metres onwards, there is a variety of relative configurations and consequently the relation between the types of distance is much more variable. This variability remains approximately constant as distances increase. Finally, note that Cost-Based distances progressively separate from Euclidean distances.

6.3 Prediction results

Figure 8 shows the final output of the whole process: the map with the prediction for each location and the map with the standard error for that prediction, which is a measure of the uncertainty for that prediction.

It is interesting to compare these results with those that arise if Euclidean distances are used. The map looks almost the same,

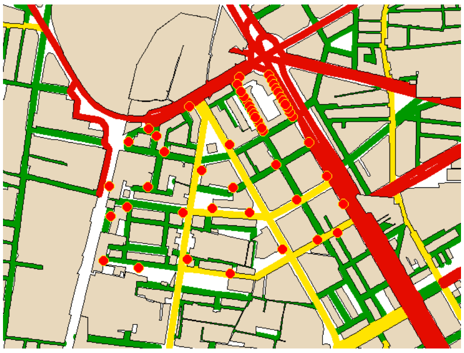


Figure 5: Initial setup. Malilla neighborhood with observation points and the classification of roads according to traffic density. Red lines represent avenues with high traffic density, while yellow and green are medium and low traffic density roads respectively. Red circles represent the locations were observations were measured.

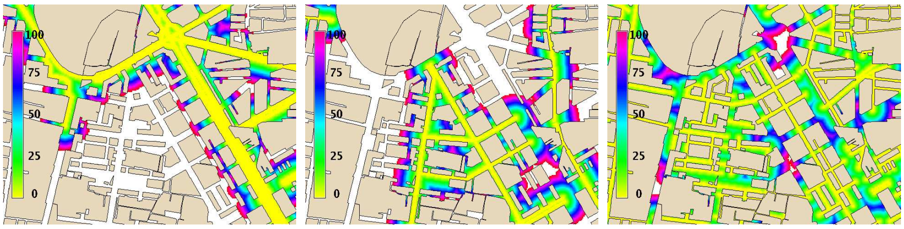


Figure 6: Distance maps to the closest road of each type.

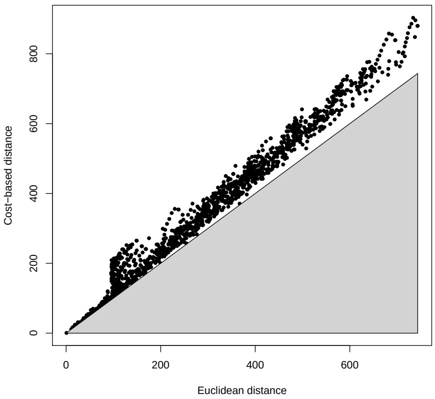
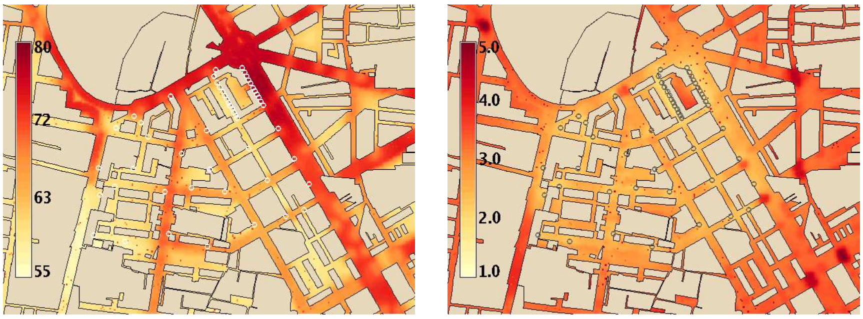


Figure 7: Cost-Based vs. Euclidean distances for all pairs of observations.



(a) noise prediction (dBA)

(b) prediction error (dBA)

Figure 8: Cost-Based kriging results. Sites of observations are lightly marked for reference.

so we should focus on the differences in the predictions and in the prediction errors.

However, note that both approaches are built over the same regression model, with Cost-Based distances to the closest roads of each type as covariates. This means that the Euclidean approach results here are not fully Euclidean. Thus, the differences that are analyzed in this section are due to the different approaches in correlation structure only.

The prediction differences between Cost-Based and Euclidean kriging for this pilot example range from -1.655 dBA to 1.876 dBA in absolute terms, and from -2.7% to 3.1% in relative terms. On average, differences are very close to zero, and in 95% of the locations, less than $\pm 1\%$.

With regard to uncertainty, the differences in standard error range from -0.25 dBA to 0.65 dBA. In relative terms, these differences span the much more relevant and wider range of -5.8% to 27.1%. In 77% of the locations, Cost-Based prediction is more accurate than Euclidean prediction, but there are a few locations where uncertainty is much higher, reaching up to 27% more error (see Fig. 9).

What is more interesting is the spatial distribution of these differences, in order to interpret in which situations and configurations the two approaches diverge (see Fig. 10).

Note that the greatest difference occurs in the enclosed area in the upper section, where the Cost-Based approach predicts a higher noise level with the greatest difference. For the Euclidean approach, the *enclose* do not exist; hence, the observations from the "quiet" road have more influence than they should because of the buildings. In contrast, the Cost-Based approach understands that the region

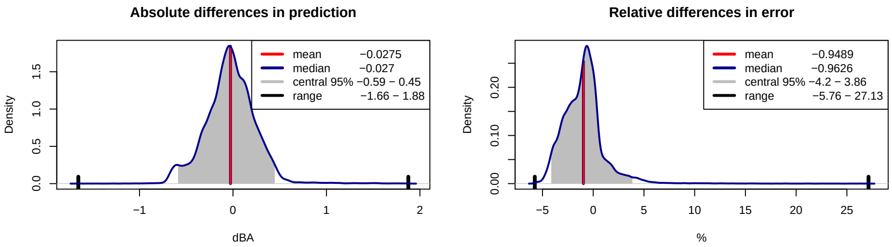
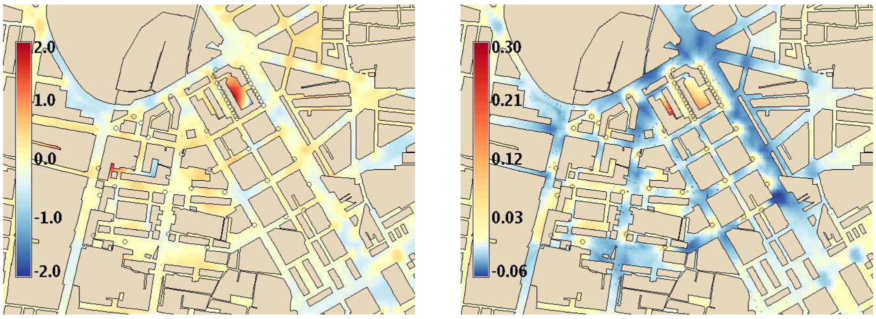


Figure 9: Distribution kernel estimates of differences between Cost-Based and Euclidean predictions and prediction errors, and sample summary values.



(a) noise prediction differences (dBA)

(b) error relative differences (%)

Figure 10: Differences between Cost-Based and Euclidean kriging results. Sites of observations are lightly marked for reference.

is much more influenced by the noise from the avenue, therefore it predicts a higher noise level.

The map with relative differences in prediction error is mostly in tones of blue, which means that the Cost-Based approach is generally more accurate. An exception is located in the same area mentioned before, which is easy to explain, since the Euclidean method ”thinks” that there are lots of observations very close around, so it assigns great precision to its prediction. On the other hand, the Cost-Based method ”knows” that observations are not that close, therefore the uncertainty is larger.

7 Conclusions

The most interesting aspect of this work lies in the general methodology for overcoming the geostatistical restriction on the homogeneity of the prediction region. Also, the combination with Geographical Information Systems enables the use of distances to relevant objects as covariates, which provide valuable information that could not be exploited otherwise.

Noise mapping in urban areas benefits from this methodology since buildings and other urban infrastructure are relevant restrictions in the noise flow. The possibility of applying geostatistical methods enables us to obtain results based on statistical models, providing reliable predictions together with estimations of uncertainty, which commonly used deterministic methods cannot provide.

References

[1] World Health Organization, Guidelines for community noise, outcome of the WHO-expert task force meeting (April 1999). URL `http://www.who.int/docstore/peh/noise/guidelines2.html`

[2] GRASS Development Team, Geographic Resources Analysis Support System (GRASS GIS) Software, Open Source Geospatial Foundation, USA (2008). URL `http://grass.osgeo.org`

[3] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2008). URL `http://www.R-project.org`

[4] R. Bivand, spgrass6: Interface between GRASS 6 geographical information system and R (2008). URL `http://grass.osgeo.org/`

[5] P. J. Ribeiro, P. J. Diggle, geoR: a package for geostatistical analysis, R-NEWS 1 (2) (2001) 14–18. URL `http://CRAN.R-project.org/doc/Rnews/`

[6] S. L. Rathbun, Spatial modelling in irregularly shaped regions: Kriging estuaries, Environmetrics 9 (1998) 109–129.

[7] L. S. Little, D. Edwards, D. E. Porter, Kriging in estuaries: As the crow flies, or as the fish swims?, Journal of Experimental Marine Biology and Ecology 213 (1997) 1–11.

[8] A. Awaida, J. Westervelt, r.cost: cumulative cost computation for GRASS GIS (2006). URL `http://grass.osgeo.org/grass63/manuals/html63_user/r.cost.html`

[9] D. G. Krige, A statistical approach to some basic mine valuation problems on the witwatersrand, Journal of the Chemical, Metallurgical and Mining Society of South Africa 52 (6) (1951) 119–139.

[10] G. Matheron, Principles of geostatistics, Economic Geology 58 (8) (1963) 1246–1266.

[11] N. A. C. Cressie, Statistics for Spatial Data, Wiley Series in Probability and Statistics, Wiley-Interscience, New York, 1993.

[12] N. Cressie, D. L. Zimmerman, On the stability of the geostatistical method, Mathematical Geology 24 (1) (1992) 45–59.

[13] P. J. Diggle, P. J. Ribeiro, Model-based Geostatistics, Springer Series in Statistics, Springer, 2007.

[14] F. C. Curriero, The use of non-euclidean distances in geostatistics, Ph.D. thesis, Department of Statistics, Kansas State University, 213 p. (1996).

[15] K. Krivoruchko, A. Gribov, Geostatistical interpolation and simulation with non-euclidean distances, in: X. Sanchez-Villa, J. Carrera, J. J. Gomez-Hernandez (Eds.), geoENV IV, Kluwer Academic Publishers, 2002, pp. 331–342.

[16] J. M. Ver Hoef, N. Cressie, R. P. Barry, Flexible spatial models for kriging and cokriging using moving averages and the fast fourier transform (fft), Journal of Computational & Graphical Statistics (2004) 265–282.

[17] D. Higdon, A process-convolution approach to modelling temperatures in the north atlantic ocean, Environmental and Ecological Statistics 5 (2) (1998) 173–190.

[18] A. Løland, G. Høst, Spatial covariance modelling in a complex coastal domain by multidimensional scaling, Environmetrics 14 (3) (2003) 307–321.

[19] A. Okabe, T. Satoh, K. Sugihara, A kernel density estimation method for networks, its computational method, and a gis-based tool, discussion paper No. 89. Unpublished (June 2008).