

# Mapping and hazard assessment of atmospheric pollution in a medium sized urban area using the Rasch model and geostatistics techniques

Francisco J. Moral<sup>a,\*</sup>, Pedro Álvarez<sup>b</sup>, José L. Canito<sup>a</sup>

<sup>a</sup>Department of Graphic Representation, University of Extremadura, Avenida de Elvas, 06071 Badajoz, Spain

<sup>b</sup>Department of Applied Economics, University of Extremadura, Avenida de Elvas, 06071 Badajoz, Spain

Received 13 June 2005; received in revised form 17 October 2005; accepted 24 October 2005

## Abstract

Researchers or decision-makers frequently need information about atmospheric pollution patterns in urbanized areas. The preparation of this type of information is a complex task, due to the influence of several individual pollutants, with different units, on the global air pollution (e.g. nitrogen dioxide concentrations, ppm, and noise, dB). In this work, a new methodology based on the formulation of the Rasch model is proposed to obtain a measure of the atmospheric pollution. Two main results were obtained after applying this method: (1) A classification of all locations according to the pollution level, which was the value of the Rasch measure; (2) The influence on the environmental deterioration of each individual pollutant (particularly, in this work, NO<sub>2</sub>, NO, CO<sub>2</sub>, CO and noise). Finally, pollution at locations where no measurements were available was estimated with the optimum interpolation technique, kriging. Kriged estimates were subsequently used to map atmospheric pollution. To illustrate the application of this two-step method (Rasch model plus interpolation), which is useful to generate hazard assessment maps based on the spatial distribution of atmospheric pollution, an example is shown.

© 2005 Elsevier Ltd. All rights reserved.

**Keywords:** Regionalized variable; Pollutant; Variogram; Kriging; GIS

## 1. Introduction

The environmental policy is an issue which attracts important attention in the European Union and, particularly, in Spain, due to the increasing alarm that economic development causes on human health and security, and the worrying events such as

the Chernobyl disaster, acid rain, greenhouse effect or destruction of the ozone layer. Today's society worry for nature and its progressive degradation, due to pollution, has as a consequence that people are demanding a less aggressive way of life for the environment, claiming clean industries, ecological produces, etc. Citizens also demand to their governing class different measures and facts to benefit the environment where they live, favouring a better life quality. Moreover, from a planning point of view, future proceedings in urban areas should consider distribution patterns of pollution.

\*Corresponding author. Tel.: +34 924 28 96 00;  
fax: +34 924 28 96 01.

E-mail address: [fjmorales@unex.es](mailto:fjmorales@unex.es) (F.J. Moral).

It is known how monitoring atmospheric pollution in urban areas involves mapping techniques that assist the decision-maker to describe and quantify the pollution at locations where no measurements were available. The preparation of pollution maps is a complex task, which is only feasible if a spatial correlation of the variable of interest is identified (e.g. Hopkins et al., 1999). And, what is more, atmospheric pollution is a very complex variable, which is affected by different chemical and physical individual pollutants.

Principal component analysis (PCA) has been used to evaluate the pollutant composition in some studies (e.g. Carlon et al., 2001). PCA is a multivariate statistical analysis converting the variables in the so-called principal components or factors, which explain the total variance of the data. Thus, the first factor explains the most variance, the second factor the next highest variance and so on. The system variance is retained by a few factors (e.g. Einax et al., 1997). With the formulation of the Rasch model as a measure technique (Álvarez and Ramiro, 1993), more information can be obtained than using PCA. Thus, a ranking of all locations, according to their level of atmospheric pollution, and the influence of each individual pollutant on the environmental deterioration for a particular area will become apparent.

The existence of a spatial correlation of atmospheric pollutants is not only a condition for an optimum interpolation of the data in space in order to generate a map of pollution, but it also provides very useful insights on the structure of the air quality patterns. Some studies have identified a strong spatial variability of air pollutants (e.g. Vardoulakis et al., 2005; Coppalle et al., 2001). The main goal of interpolation is to discern the spatial patterns of atmospheric pollution concentrations by estimating values at unsampled locations based on measurements at sample points. Geostatistics provides an advanced methodology to quantify the spatial features of the studied variables and enables spatial interpolation, kriging (e.g. Isaaks and Srivastava, 1989; Goovaerts, 1997). In addition, geographical information systems (GIS) and geostatistics have opened up new ways to study and analyse spatial distributions of regionalized variables, i.e. distributed continuously on space (e.g. McGrath et al., 2004; Korre et al., 2002). Moreover, they have become useful tools for the study of hazard assessment and spatial uncertainty (e.g. Goovaerts, 2001). Without a GIS, analysis and

management of large spatial databases may not be possible.

Many air pollution studies have employed distance-weighting methods (e.g. Phillips et al., 1997), but kriging is the only one which incorporates the spatial correlation into its estimation algorithm. Kriging has been used more widely (e.g. Tayanc, 2000; McGrath et al., 2004) due to its many advantages (Goovaerts, 1997). Although kriging requires an abundance of sample points to be an accurate spatial interpolation method (e.g. Myers, 1991), even when relatively small data sets and not exhaustive samplings are available it is a reliable technique for investigating the distribution and sources of pollutants (e.g. Carlon et al., 2001).

In this work, the spatial distribution of atmospheric pollution was analysed for an urban area. A new methodology was proposed, which consisted of two phases: first, the Rasch model was chosen to establish the influence of each individual pollutant and obtain a global measure of pollution; secondly, a GIS and geostatistical techniques were used to reveal the spatial distribution patterns of a regionalized variable, atmospheric pollution, and provide a basis for hazard assessment.

## 2. Materials and methods

### 2.1. The data set

Data were collected at 60 locations (sample points) in the city of Badajoz, southwestern Spain. Principally, they were located in the centre of the urban area. The population of this city is around 130 000 people. The main source of air pollution is the traffic of vehicles because industry is not important around or in the city. At each sample point, four chemical pollutants were measured, CO, SO<sub>2</sub>, NO and NO<sub>2</sub>, with a mobile monitoring unit equipped with real-time analysers, Metrosonic pm-770, which generated a mean value of each pollutant for a interval of 15 min, and noise was measured with a digital sonometer, Cel-256, considering intervals of 15 min for each measurement too. Next, it was necessary to obtain a measurement of the atmospheric pollution using different individual pollutants with different units, ppm for the chemical pollutants and dB for the noise. This important problem may be solved with the application of a measure technique based on the Rasch model (Álvarez, 2005).

All individual pollutants were measured at each location considering seven hourly intervals. Therefore, seven measurements of each pollutant were obtained at each sample point but, in this work, only the maximum value was considered, independent of the hour in which the measurement was conducted.

It is convenient to indicate that the four chemical pollutants and the noise were selected to illustrate the methodology we propose. Obviously, the consideration of other additional pollutants could improve the final estimate of atmospheric pollution, but the method would be the same.

## 2.2. The Rasch model

While the Rasch model is well known for its efficiency and precision of transforming categorical item responses to objective scale measures, it also has an interesting capacity to consolidate data that are already reported sometimes in several scale metrics. If guided by a reasonably coherent conceptual goal, the Rasch model can synthesize and consolidate seemingly disparate data into a uniform analytical framework. The purpose of this procedure is to transcend several heterogeneous physical measures and consolidate them into an overall variable that simplifies interpretation of air pollution exposure.

A key characteristic of the Rasch model is the transformation of raw data to linear units that operationally define a latent variable or theoretical construct. This variable is the amalgamation of noncategorical measures that are conceptually related to a hypothesized latent trait. Their unrelated independent units are then categorized with uniform rating scales and transformed to common logit units with Rasch measurements. However, a question can arise about their meaningfulness. If several agents were originally measured in units that fundamentally differ in nature, it is not immediately obvious how to interpret them as ratings (one would legitimately think this situation is similar to summing apples and oranges). Fortunately, the meaningfulness of these measures is derived by their probabilistic relations to an overriding theoretical construct, their empirical convergence on an invariant, unidimensional structure. If agents have an a priori conceptual relationship with an abstract, hypothetical construct, then their empirical reformulation as ordered categories frees these agents from their prior metric constraints. By describing

agents in terms of uniform rating categories, so that high agent values would be equivalent to the highest rating category, and low values to the lowest rating category or level, unrelated agents and dimensions acquire common ordinality (intermediate categorical values could be obtained by interpolation). Through this numerical manipulation, independent scale quantities can be expressed as common ratings ranging from low to high. The rationale for changing several noncategorical, continuous measures to ordered rating categories is a fundamental conviction that these measures are related to an important overall construct, and a desire to better understand their inter-relations on this construct.

Let  $n$  be the different locations in the town of Badajoz where measurements of each pollutant,  $i$ , were carried out. We define a latent variable or construct, atmospheric pollution,  $X_{ni}$ , in which  $n$  refers to the location where the measurement is conducted and  $i$  refers to the pollutant. In the case we are studying,  $\beta_n$  ( $n = 1, 2, \dots, 60$ ) refers to the 60 locations (Plaza Dragones, Venero, Puente Viejo, etc.; column 'Location' in Table 2) where the measurements were carried out, and  $\delta_i$  ( $i = 1, 2, 3, 4$ , or 5) refers to the five individual pollutants (noise -1-, CO -2-, SO<sub>2</sub> -3-, NO -4- and NO<sub>2</sub> -5-). For instance,  $X_{39,2}$  means the measurement of the pollutant  $i = 2$  (CO) at the location  $n = 39$  (Isidro Pacense).

$X_{ni}$ , like any another latent variable, can be regarded as a straight line along which items (pollutants),  $\delta_i$ , and locations,  $\beta_n$ , are located. The line ranges from less atmospheric pollution to more for any urban location and is operationally defined by the five pollutants previously indicated. The further to the right a sample point is located, the greater its pollution. A way to establish the appropriate placement of the locations along the line in terms of items, representing simultaneously the pollution of the sample points with respect to the pollutants and vice versa, is as follows: for example,  $X_{01}$ ,  $X_{02}$ ,  $X_{03}$  and  $X_{04}$ , means that pollutants  $\delta_1$ ,  $\delta_2$ ,  $\delta_3$  and  $\delta_4$  have been measured at the location  $\beta_0$ . In this framework, any pollutant has some probability of appearing at any location, and the measurement problem is to represent their linear differences on a probabilistically additive, equal interval scale. The numerical gradient for this scale is called a logit (log odds) and established by estimating ordered category threshold parameters for ratings of pollutant measures collected at locations. To estimate pollutant and location

positions, this approach was formally implemented in a Rasch model for rating scales (Andrich, 1988; Wright and Masters, 1982).

If at a location  $\beta_n$  all the pollutants  $\delta_i$  are detected, then  $\beta_n$  would be placed to the right of these items  $\delta_i$ . On the contrary, if the pollutants are not detected, then  $\beta_n$  will be located on the left of all  $\delta_i$ . Diagram 1 in Fig. 1 illustrates how the location  $\beta_0$ , and the pollutants  $\delta_1, \delta_2, \delta_3$ , and  $\delta_4$  are located along the line that represents atmospheric pollution,  $X_{ni}$ . In this case, pollutants  $\delta_1, \delta_2$ , and  $\delta_3$ , are closer to the left end of the line than location  $\beta_0$  and pollutant  $\delta_4$ ; consequently, at that location, there is pollution due to those three pollutants,  $\delta_1, \delta_2$ , and  $\delta_3$ , but not to  $\delta_4$ . In diagram 2 (Fig. 1), location  $\beta_1$  would have not atmospheric pollution; it is not affected by any pollutant. In diagram 3 (Fig. 1), location  $\beta_2$  would have atmospheric pollution; it is affected by all pollutants. If there are two or more locations, their difference in terms of pollution would be given by their relative positions with respect to the number of pollutants. Thus, the latent variable atmospheric pollution,  $X_{ni}$ , is the continuum, represented on a line, along which are located the parameters  $\delta_i$  for the pollutants and  $\beta_n$  for the locations. In diagram 4 (Fig. 1), location  $\beta_3$  surpasses no pollutant; location  $\beta_4$  only surpasses pollutant  $\delta_1$ ; location  $\beta_5$  surpasses pollutants  $\delta_1$  and  $\delta_2$ ; and location  $\beta_6$  surpasses all three pollutants. Therefore,  $\beta_3$  is the location with least atmospheric pollution, and  $\beta_6$  has the most. Pollutant  $\delta_1$  does not

affect the location  $\beta_3$  and affects locations  $\beta_4, \beta_5$ , and  $\beta_6$ . For pollutant  $\delta_2$  is the following; that does not affect the locations  $\beta_3$  and  $\beta_4$ , and affects locations  $\beta_5$  and  $\beta_6$ . Finally, pollutant  $\delta_3$  does not affect locations  $\beta_3, \beta_4$ , and  $\beta_5$ , and affects location  $\beta_6$ . In this case,  $\beta_6$  is the most polluted location since it is affected by all the pollutants,  $\delta_1, \delta_2$ , and  $\delta_3$ ;  $\beta_3$  is the least polluted location since it is not affected by any pollutant. On the other hand,  $\delta_1$  is the most frequent pollutant, in the sense that it affects more locations than any other pollutant. Therefore,  $\delta_1$  is a more frequent pollutant than  $\delta_2$ , and this is more frequent than  $\delta_3$ .

Consider the dichotomous variable atmospheric pollution,  $X_{ni}$ , which describes the fact that a location  $\beta_n$  is affected by the pollutant  $\delta_i$ . If  $X_{ni} = 1$  then location  $\beta_n$  is said to be affected by that pollutant, and if  $X_{ni} = 0$  then location  $\beta_n$  is said not to be affected by that pollutant. One way of relating the positions of the locations  $\beta_n$  and of the pollutant  $\delta_i$  with the dichotomous variable in terms of probability is:

If  $(\beta_n - \delta_i) > 0$ , means that  $\beta_n$  is on the right of  $\delta_i$  in the line where  $X_{ni}$  is defined; then the probability that the location  $\beta_n$  is affected by the pollutant  $\delta_i$  is higher than 0.5.

If  $(\beta_n - \delta_i) < 0$ , means that  $\beta_n$  is on the left of  $\delta_i$  in the line where  $X_{ni}$  is defined; then the probability that the location  $\beta_n$  is affected by the pollutant  $\delta_i$  is lower than 0.5.

If  $\beta_n = \delta_i$ , means that  $\beta_n$  and  $\delta_i$  coincide in the line where  $X_{ni}$  is defined; then the probability that the location  $\beta_n$  is affected by the pollutant  $\delta_i$  is 0.5,

that is,

If  $(\beta_n - \delta_i) > 0$ , then  $P[X_{ni} = 1] > 0.5$ .

If  $(\beta_n - \delta_i) < 0$ , then  $P[X_{ni} = 1] < 0.5$ .

If  $\beta_n = \delta_i$ , then  $P[X_{ni} = 1] = 0.5$ .

The difference  $(\beta_n - \delta_i)$  can range from  $-\infty$  to  $+\infty$ , and the probability from 0 to 1, i.e.

$$-\infty \leq (\beta_n - \delta_i) \leq +\infty \text{ and } 0 \leq P[X_{ni} = 1] \leq 1.$$

If we use the difference as an exponent of e, then:

$$0 \leq e^{(\beta_n - \delta_i)} \leq +\infty.$$

With a further adjustment, we can bring the expression into the interval from zero to one:

$$0 \leq \left\{ \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \right\} \leq 1.$$

Diagram 1

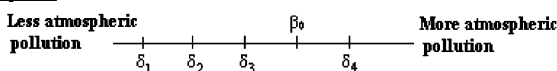


Diagram 2

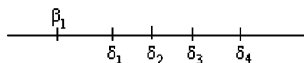


Diagram 3

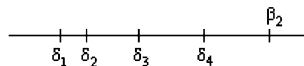


Diagram 4

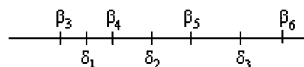


Fig. 1. Representation of the latent variable, atmospheric pollution, as a straight line.  $\beta_n$  is the location  $n$ ;  $\delta_i$  is the pollutant  $i$ . Diagram 1 illustrates the case of a location  $\beta_0$  which is affected by pollutants  $\delta_1, \delta_2, \delta_3$ , but not by  $\delta_4$ . In diagram 2, location  $\beta_1$  is not affected by any pollutant. In diagram 3, location  $\beta_2$  is affected by all pollutants. Diagram 4 shows a generalization for some locations and pollutants;  $\beta_3$  is not affected by any pollutant;  $\beta_4$  is affected by the pollutant  $\delta_1$ ;  $\beta_5$  is affected by the pollutants  $\delta_1$  and  $\delta_2$ ;  $\beta_6$  is affected by all pollutants,  $\delta_1, \delta_2$  and  $\delta_3$ .

The relationship can be written as (Álvarez and Pulgarín, 1996)

$$P[X_{ni} = 1; \beta_n, \delta_i] = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}},$$

which is the probability that location  $n$  has the pollutant for item  $i$ , given the parameters  $\beta_n$  and  $\delta_i$ . This is the formula obtained by Rasch (1980) in his treatise on latent variables.

### 2.3. Geostatistics

The formulation of the Rasch model allowed one to obtain values of the atmospheric pollution for all sample points, incorporating information of five individual pollutants. Later, it was necessary to estimate the atmospheric pollution at other locations where direct measurements were not carried out. Since the factors that determine the values of environmental variables are numerous, largely unknown in detail, and interact with a complexity that we cannot unravel, we can regard their outcomes as random. If a stochastic point of view is adopted, then there is not just one value for a property but a whole set of values at each point in space. We regard the observed value there as one drawn at random according to some law, from some probability distribution. This point of view, when the studied variable (atmospheric pollution) is considered random and distributed continuously on the experimental area (regionalized variable), is adopted to use geostatistics as an estimation technique.

Geostatistics can be defined as the set of tools and techniques to analyse the spatial patterns and predict at unsampled locations the values of a continuous variable distributed in space or in time. It is also denominated spatial statistics (e.g. Goovaerts, 1997).

In this study, three phases were completed to conduct the geostatistical work (e.g. Isaaks and Srivastava, 1989):

- (1) *Exploratory analysis of data*: Data were studied without considering their geographical distribution. Statistics was applied to check data consistency, removing outliers and identifying statistical distribution where data came from.
- (2) *Structural analysis of data*: Spatial distribution of the variable was analysed. Spatial correlation or dependence can be quantified with semivariograms, or simply variograms, which also can characterize and determine distributions patterns such as aggregation, randomness, unifor-

mity and spatial trend. Variogram function relates the semivariance, half the expected squared difference between paired data values  $Z(x_i)$  and  $Z(x_i + h)$ , to the lag distance,  $h$ , by which sample points are separated. For discrete sampling locations, the function is estimated as

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \{Z(x_i) - Z(x_i + h)\}^2,$$

where  $\gamma(h)$  is the experimental semivariance value at distance interval  $h$ ,  $Z(x_i)$  are the measured sample values at sample points  $x_i$ , in which there are data at  $x_i$  and  $x_i + h$ ;  $N(h)$  is the total number of sample pairs within the distance interval  $h$ . The variogram shows the degradation of spatial correlation between two points of space when the separation distance increases. This function has two components: (a) The nugget effect, which characterizes the discontinuity jump observed at the origin of distances, quantifies the short-term, erratic variations of the studied phenomenon plus measurements and data errors. (b) The increasing part of the variogram, which may reach the sill (theoretical sample variance), leveling off the curve, for a distance called range, or keep on increasing continuously with distance. The non-nugget part of the variogram measures the nonrandom part of the phenomenon and models its average medium-scale behaviour in space.

When an experimental variogram is defined, i.e. some points of a variogram plot are determined by calculating variogram at different lags, a model (theoretical variogram) should be fitted to the points. Although there are some statistical techniques to justify the choice of a theoretical variogram (e.g. Cressie, 1985), subjective criteria and previous experiences are the main tools to choose one.

- (3) *Predictions*: The main objective of a geostatistical study is to obtain estimates of values of the studied variable at unsampled locations, considering the spatial distribution pattern and integrating information from sample points and observed or known trends, if they exist. Geostatistics offers a great variety of methods that provide estimates for unsampled locations. These methods are known as kriging, in honor of Danie Krige, who first formulated this form of interpolation in 1951. Kriging is regarded as the best linear unbiased estimator (BLUE). Weights for sample values are calculated based

on the parameters of the variogram model. The sum of all weights must be one due to the necessity for ensuring that estimates are unbiased. Moreover, kriging variances or estimation errors need to be minimized.

All different types of kriging are distinguished depending on the chosen model for the trend of the random function. In this work, the geostatistical interpolation method known as ordinary kriging was used (e.g. Goovaerts, 1997).

2.4. Data treatment

Different software packages were used to analyse raw data. Winsteps 3.35 computer program was employed to conduct the formulation of the Rasch model (Linacre, 2000). The geostatistical analysis, including all three phases described in the previous section, were carried out with the extension Geostatistical Analyst of the GIS software ArcGIS (version 8.3).

Maps of kriged estimates provided a visual representation of the distribution of the atmospheric pollution in Badajoz. These maps were produced with the ArcMap module of the ArcGIS, after conducting the geostatistical study.

3. Results and discussion

3.1. Determination of atmospheric pollution at sample points

The formulation of the Rasch model to obtain a measure of atmospheric pollution at each sample

point, which would take into account the different contribution of five pollutants (NO<sub>2</sub>, NO, CO<sub>2</sub>, CO and noise), was achieved through the stages shown in Fig. 2.

The first stage was a previous transformation of the data to a common scale (Wright and Masters, 1982). Pollutant measures were categorically coded according to a plan where each location was rated on a scale (0–5) for each pollutant. Table 1 presents the assignment of categorical values across pollutant measures. As it was indicated in Section 2.2, the minimum and maximum values of the scale were assigned to the maximum and minimum values of each pollutant. Other intermediate values for all different pollutants were obtained through interpolation. For example, at Hotel Río location ( $\beta_n = 20$ , Table 2), the measurements of all pollutants were: SO<sub>2</sub> = 0.2 ppm; NO<sub>2</sub> = 0.28 ppm; NO = 1.81 ppm; CO = 20.7 ppm; noise = 85.34 dB. After coding them, using the categories shown in Table 1, the rating scale values were: SO<sub>2</sub> → 2; NO<sub>2</sub> → 3; NO → 4; CO → 3; noise → 4.

After processing all data, results shown in Table 2 were obtained. The raw score and measure values should be highlighted. The first one shows the sum of points of all scores (rating scale categories) for each pollutant. The second one indicates the measures for the locations according to their raw score. For example, at Hotel Río location, raw score = 2 + 3 + 4 + 3 + 4 = 16, i.e. the rating scale values for all pollutants were summed. The measure value is estimated with the Winsteps program, where the approach previously described in Section 2.2 was implemented and parameters  $\delta_i$  and  $\beta_n$  are located in the straight line which represents the

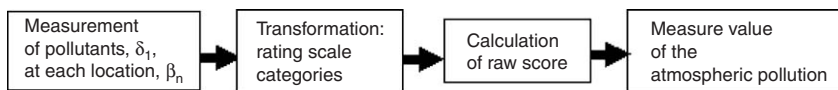


Fig. 2. Schematic diagram of the stages involved in the formulation of the Rasch model.

Table 1  
Pollutants measures recoded into rating scale categories

| SO <sub>2</sub> (ppm) | NO <sub>2</sub> (ppm) | NO (ppm)  | CO (ppm)  | Noise (dB)  | Rating scale value |
|-----------------------|-----------------------|-----------|-----------|-------------|--------------------|
| 0                     | 0                     | 0         | 0         | 50–58.33    | 0                  |
| 0.1                   | 0–0.12                | 0–0.56    | 0–8.4     | 58.33–66.66 | 1                  |
| 0.2                   | 0.12–0.24             | 0.56–1.12 | 8.4–16.8  | 66.66–74.99 | 2                  |
| 0.3                   | 0.24–0.36             | 1.12–1.68 | 16.8–25.2 | 74.99–83.32 | 3                  |
| 0.4                   | 0.36–0.48             | 1.68–2.24 | 25.2–33.6 | 83.32–91.65 | 4                  |
| 0.5                   | 0.48–0.60             | 2.24–2.80 | 33.6–42   | 91.65–100   | 5                  |

Table 2

Results obtained after applying the Rasch model: sum of points of the common scale for all individual pollutants (raw score) and atmospheric pollution level (measure)

| Number ( $\beta_n$ ) | Location          | Raw score | Measure |
|----------------------|-------------------|-----------|---------|
| 51                   | Plaza Dragones    | 24        | 74.9    |
| 8                    | Venero            | 22        | 68.4    |
| 28                   | Puente Viejo      | 20        | 64.5    |
| 34                   | Plaza Minayo      | 19        | 62.8    |
| 43                   | Puente S. Roque   | 19        | 62.8    |
| 9                    | Cruce Olivenza    | 18        | 61.2    |
| 52                   | Plz. Constitución | 18        | 61.2    |
| 18                   | Cruce Sevilla     | 17        | 59.7    |
| 20                   | Hotel Río         | 16        | 58.1    |
| —                    | —                 | —         | —       |
| 36                   | Plaza Soledad     | 9         | 45.6    |
| 39                   | Isidro Pacense    | 9         | 45.6    |
| 54                   | Fco. Luján        | 9         | 45.6    |
| 1                    | E. II.II.         | 8         | 43.3    |
| 4                    | Avda. Sinfor.     | 8         | 43.3    |
| 44                   | Parque Legión     | 8         | 43.3    |
| 45                   | Plaza Huelva      | 8         | 43.3    |
| 47                   | Antonio Cuéllar   | 8         | 43.3    |
| 56                   | Plaza Conquistad. | 8         | 43.3    |

Only the first and last locations of the list are shown. In total there are 60 locations.

Table 3

Influence of each individual pollutant on the atmospheric pollution in Badajoz

| Pollutant       | Raw score | Measure |
|-----------------|-----------|---------|
| Noise           | 175       | 114.5   |
| NO <sub>2</sub> | 164       | 131.3   |
| NO              | 153       | 139.8   |
| CO              | 125       | 152.3   |
| SO <sub>2</sub> | 123       | 175.4   |

The raw score is the sum of points of the common scale for each pollutant considering all locations (60). The measure indicates the position of each pollutant along the straight line that represents the latent variable, atmospheric pollution.

latent variable, using an unconditional maximum likelihood procedure (Wright and Panchapakesan, 1969). In Table 2, all locations were arranged in measure order, obtained by the formulation of the Rasch model.

Another interesting result, which was obtained after processing the data, is shown in Table 3, where the influence of each individual pollutant on the atmospheric pollution can be observed. Thus, noise is the most influential pollutant on environmental deterioration in Badajoz, since it obtained the

highest raw score, and will correspond to the lowest measure, i.e. all locations are affected by noise. Unlike noise, sulphur dioxide, SO<sub>2</sub>, is the pollutant with a lowest raw score but the highest measure; therefore, the influence on the environmental deterioration in Badajoz is the least important among all individual pollutants, i.e. SO<sub>2</sub> does not affect as much as noise.

The Rasch model can provide more information through the misfitting analysis (Álvarez, 2005). However, the study of misfits has not been included in this paper because we were interested in the global atmospheric pollution.

### 3.2. Spatial distribution maps of atmospheric pollution

Some geostatistical tools were required to estimate at any unsampled point, using as previous information all measure values of atmospheric pollution at sample sites (Table 2). During the exploratory analysis of data, it was revealed that they were distributed lognormally. The histogram of the measures showed a tail of high values to the right, making the median (51.5) less than the mean (51.75). The coefficient of skewness was 1.23 and the kurtosis was 4.93. After taking logarithms, data showed an appearance almost normal (median = mean = 3.94) and, consequently, the skewness was lower, 0.83; the kurtosis was also lower, 3.8. A normal QQ plot, a graph of the quantiles of the input dataset versus quantiles of the standard normal distribution (Fig. 3), confirms that data are lognormally distributed. Thus, transformed data were used for the following geostatistical analyses.

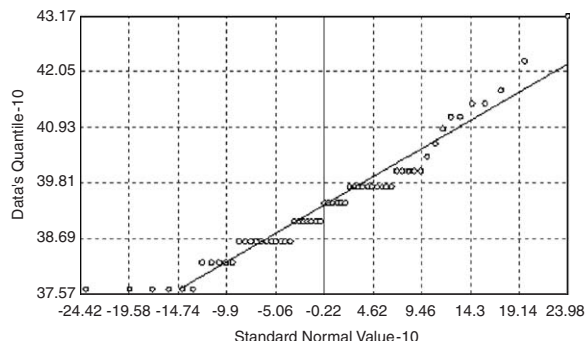


Fig. 3. Normal QQ plot of the data, previously transformed taking logarithms. In general, the points lie close to the straight line which indicates perfect normality. The main departure from this line occurs at high or low values.



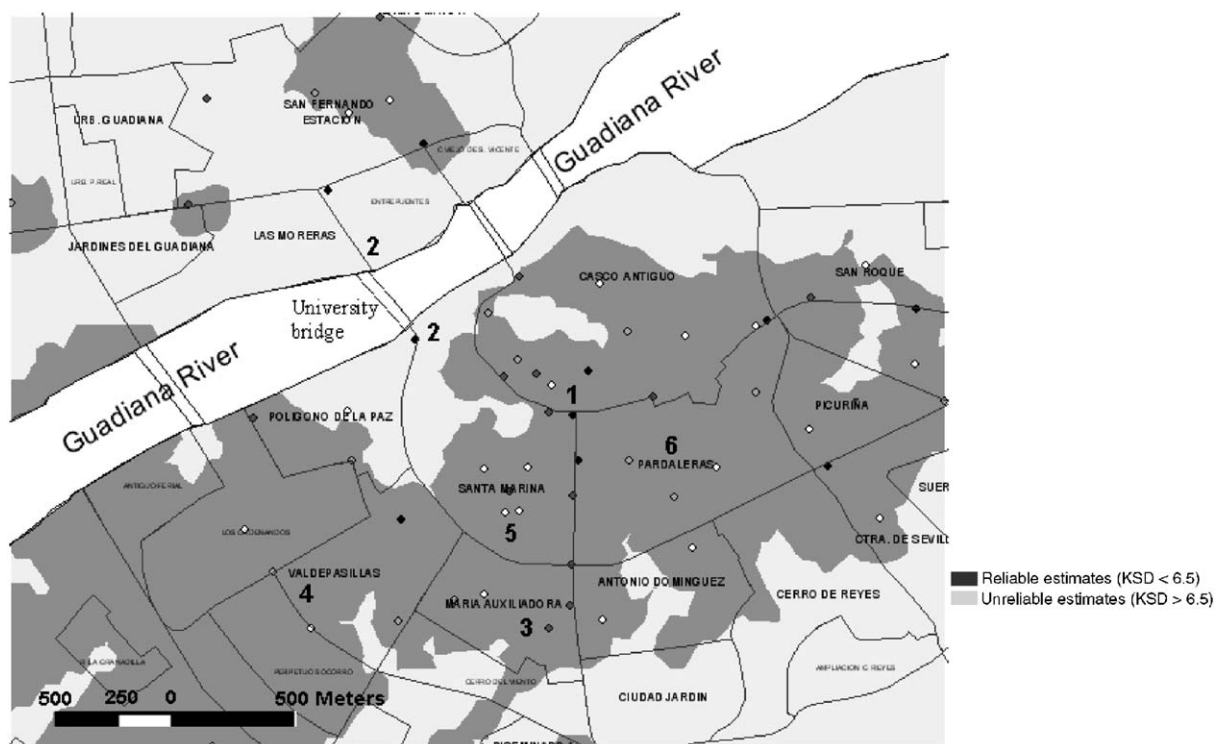


Fig. 5. Map of reliable (kriging standard deviation, KSD, below 6.5) and unreliable (kriging standard deviation, KSD, above 6.5) estimates of atmospheric pollution in the centre of Badajoz. Circles represent sample points.

map, based on the spatial distribution of atmospheric pollution, is shown in Fig. 4. Extreme pollution areas correspond with locations where traffic is intense and with many stops because of the existence of numerous traffic lights, for example, in Casco Antiguo (area 1, Fig. 4) or around the University bridge (area 2, Fig. 4). Other areas where traffic is intense but with roundabouts, for instance, María Auxiliadora (area 3, Fig. 4) or Valdepasillas (area 4, Fig. 4), are less polluted.

One of the strengths of using the geostatistical methods is that it is possible to calculate a statistical measure of the reliability of the maps of estimates. Thus, another output of kriging is the kriging variance, or its square root, the kriging standard deviation (KSD), which is calculated for each sample point. KSD is related to the sample distribution and variogram structure. KSD can be mapped similarly to estimates. These maps give an idea of the quality of the estimates at different places. As Webster and Oliver (2001) indicated, the maps of the estimation variance or standard deviation should be used with caution. The reliability of kriging depends on how accurately the variation is represented by the chosen spatial model.

Our estimates could be more reliable than they appear to be if the nugget effect is overestimated. In this work, the nugget effect was high, as it was previously mentioned, so we can consider that predictions are, at least, as reliable as the value of the KSD indicates.

If we consider that KSD lower than 6.5 is acceptable, a map with only two classes, one above 6.5 and another below this value, can be used to establish locations where predictions are reliable. In general, areas with many sample points, e.g. Santa Marina (area 5, Fig. 5) or Casco Antiguo (area 1, Fig. 5), or areas where data were sparse but evenly distributed, e.g. Pardaleras (area 6, Fig. 5) or Valdepasillas (area 4, Fig. 5), had the most reliable estimates.

Based on the combination of kriging map and KSD map, the probability map (e.g. Goovaerts, 1997) of atmospheric pollution  $> 52$  (Rasch measure) was produced (Fig. 6). This threshold, 52, was chosen because above it, pollution was high or extreme according to the classification we considered previously. This map shows that the areas with high risk of atmospheric pollution are at the convergence of Santa Marina (area 5, Fig. 6),

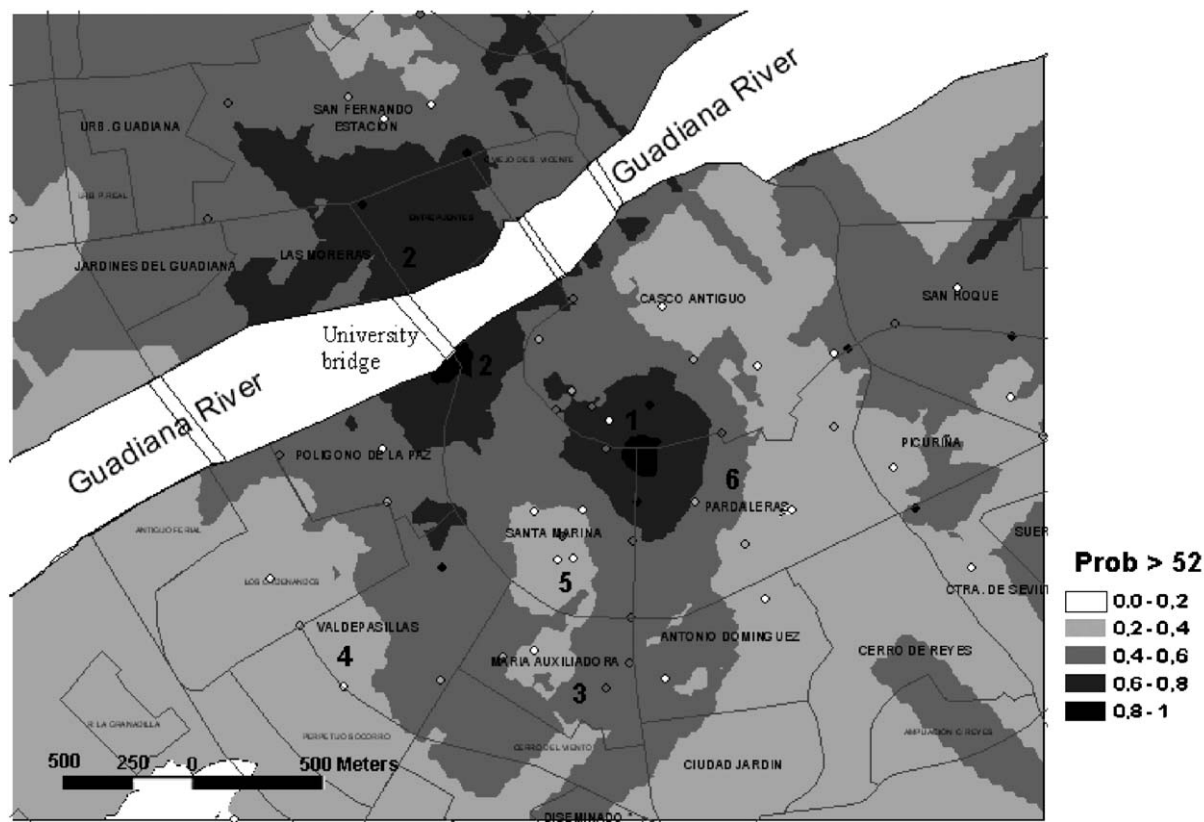


Fig. 6. Probability map of atmospheric pollution > 52 (Rasch measure) in the centre of Badajoz. Circles represent sample points.

Pardaleras (area 6, Fig. 6) and Casco Antiguo (area 1, Fig. 6), and around the University bridge (area 2, Fig. 6), the same that showed the highest levels of pollution according to classification of Fig. 4. Areas with high probabilities, for instance, above 0.6, may be regarded as dangerous, where atmospheric pollution is likely to be higher than 52. Conversely, areas with low probabilities, for example, below 0.4, may be regarded as safe, because atmospheric pollution is less likely to be higher than 52. In this type of map, the probabilities provide a measurement of confidence for hazard assessment of atmospheric pollution.

#### 4. Conclusions

The formulation of the Rasch model has been proposed to define a measure of atmospheric pollution which integrates different measurements of individual pollutants: CO, SO<sub>2</sub>, NO<sub>2</sub>, NO and noise. Moreover, this method can detect the influence of each pollutant on the environmental

deterioration and the pollution measurement for each location.

Later, geostatistical techniques were used to estimate atmospheric pollution throughout the experimental area. After analysing data and obtaining a good variogram, kriging was used to estimate at unsampled locations. Subsequently, Kriged estimates were employed to map atmospheric pollution. Useful information for hazard assessment was also obtained when a probability map, based on kriging interpolation and KSD, was produced.

The combination of the Rasch model and geostatistical techniques is a powerful tool to develop an appropriate environmental and managing policy.

#### References

- Álvarez, P., 2005. Several Noncategorical Measures Define Air Pollution Construct. Rasch Measurement in Health Science. JAM Press, Maple Grove, MN, USA.
- Álvarez, P., Pulgarín, A., 1996. The Rasch Model. Measuring the impact of scientific journals: Analytical Chemistry. Journal of the American Society for Information Science 47 (6), 458–467.

- Álvarez, P., Ramiro, A., 1993. Measuring pollution in Badajoz. In: Second Conference on Statistics, Earth and Space Sciences. CHESM-93: Chemometrics and Environmetrics Meeting, Satellite, Bologna, Italy.
- Andrich, D., 1988. Rasch Model for Measurement. Sage Publications, Newbury Park, CA, USA.
- Carlson, C., Critto, A., Marcomini, A., Nathanail, P., 2001. Risk based characterisation of contaminated industrial site using multivariate and geostatistical tools. *Environmental Pollution* 111, 417–427.
- Coppalle, A., Delmas, V., Bobbia, M., 2001. Variability of NO<sub>x</sub> and NO<sub>2</sub> concentrations observed at pedestrian level in the city centre of a medium sized urban area. *Atmospheric Environment* 35, 5361–5369.
- Cressie, N., 1985. Fitting variogram models by weighted least squares. *Mathematical Geology* 17 (5), 563–586.
- Einax, J.W., Zwanzinger, H.W., Geib, S., 1997. Chemometrics in Environmental Analysis. VCH A Wiley Company, Weinheim.
- Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York.
- Goovaerts, P., 2001. Geostatistical modeling of uncertainty in soil science. *Geoderma* 103, 3–26.
- Hopkins, L.P., Ensor, K.B., Rifai, H.S., 1999. Empirical evaluation of ambient ozone interpolation procedures to support exposure models. *Journal of the Air & Waste Management Association* 49, 839–846.
- Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York.
- Korre, A., Durucan, S., Koutroumani, A., 2002. Quantitative-spatial assessment of the risks associated with high Pb loads in soils around Lavrio, Greece. *Applied Geochemistry* 17, 1029–1045.
- Linacre, J.M., 2000. Winsteps (Computer Program and Manual). MESA Press, Chicago.
- McGrath, D., Zhang, C.S., Carton, O.T., 2004. Geostatistical analyses and hazard assessment on soil lead in Silvermines area, Ireland. *Environmental Pollution* 127, 239–248.
- Myers, D.E., 1991. Interpolation and estimation with spatially located data. *Chemometrics and Intelligent Laboratory Systems* 11, 209–228.
- Phillips, D.L., Tingey, D.T., Lee, E.H., Herstrom, A.A., Hogsett, W.E., 1997. Use of auxiliary data for spatial interpolation of ozone exposure in southeastern forests. *Environmetrics* 8 (1), 43–61.
- Rasch, G., 1980. Probabilistic Models for Some Intelligence and Attainment Tests (1960). University of Chicago Press, Denmark (Revised and expanded ed.).
- Tayanc, M., 2000. An assessment of spatial and temporal variation of sulfur dioxide levels over Istanbul, Turkey. *Environmental Pollution* 107, 61–69.
- Vardoulakis, S., Gonzalez-Flesca, N., Fisher, B.E.A., Pericleous, K., 2005. Spatial variability of air pollution in the vicinity of a permanent monitoring station in central Paris. *Atmospheric Environment* 39, 2725–2736.
- Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists. Wiley, Chichester.
- Wright, B.D., Masters, G.N., 1982. Rating Scale Analysis. MESA Press, Chicago.
- Wright, B.D., Panchapakesan, N., 1969. A procedure for sample-item analysis. *Educational and Psychological Measurement* 29, 23–48.