DESCRIBING THE TOTAL NUMBER OF DIAGNOSED CASES OF AIDS BY MEANS OF GEOSTATISTICS

Sueli Aparecida MINGOTI[1]
Arlene Guimarães LEITE[1]
Gilmar ROSA[1]

- ABSTRACT**:** In this paper a geostatistical model is fitted using the total number of diagnosed cases of AIDS in the 1996-1999 period as observed in the municipal districts of Minas Gerais state in Brazil. The prediction model is used to estimate the number of AIDS cases for municipal districts in Minas Gerais which did not present any reported cases in the period of study. It will be shown that there was no great difference between the Mingoti and Pantuzzo's (1998) semi-variogram model estimated by using the 1990-1995 AIDS data and the semi-variogram model estimated by using the 1996-1999 AIDS data. The discussion presented in this paper indicates that the geostatistical model has a good performance in describing the number of diagnosed cases of AIDS in Minas Gerais state.

- KEYWORDS: Geostatistics; semi-variogram; prediction, AIDS.

1 Introduction

Nowdays, several aspects of the AIDS disease have being studied by researchers worldwide. Many statistical models for prediction of the number of new cases of AIDS have been investigated. The more common statistical models are based in extrapolation (McEvoy and Tillet, 1985; Morgan and Curran, 1986) or back-calculation (Brookmeyer and Gail, 1988). The extrapolation models are based on the fitting of a curve to the AIDS incidence observed in a certain population of interest, considering the time dependence (Jager and Ruitenberg, 1988). These curves are fitted using the regression methodology and projection of the number of cases of the disease is performed for future years that do not belong to the period corresponding to that observed sample. Therefore, the fitted regression model is used for projections for years outside the time domain of the observed data. Larger errors can be involved in this procedure because the true curve for future years may be different from the curve fitted by using the observed data, especially for a long projection period. The back-calculation is a method for short term projection and depends on the estimates of the number of individuals previoulsly infected with the AIDS in the target population. The difficulty for the implementation of this method is its dependence on the information about the probability distribution of the incubation time of the disease. This probability distribution may change over the years due to the increase of

[1]Departamento de Estatística, Universidade Federal de Minas Gerais – UFMG, CEP 30161-970, Belo Horizonte, MG, Brasil, E-mail: sueli@est.ufmg.br

scientific knowledge about the factors associated with the transmission of the HIV virus and, consequently, the development of new medications and new policies for prevention of the disease. Back-calculation models which allow the updating of this probability distribution are presented by Brookmeyer and Liao (1990) and by Solomon and Wilson (1990). Other back-calculation models which incorporate the information about the patient's age when infected by HIV are presented in Rosenberg (1995) and in Becker and Marschner (1993). A very interesting reference about the statistical models used to describe the AIDS dissemination dynamics is Brookmeyer (1996). Other references are Wu and Ding (1999), Mayer-Hamblett and Self (2001), Ding and Wu (2000), FoulkEs and Gruttola (2002), Faucett et. al. (2002).

Although the number of diagnosed cases varies between cities or areas in the same state or country, spatial information is usually ignored in these statistical models. However, considering that the dynamics of people's migration can also contribute to the dissemination of the disease due to factors that are associated with the transmission of the HIV virus, such as sexual contact with infected people or contact with contamined blood, spatial information becomes an important factor in the structure of statistical models for prediction (Caiffa, Mingoti et. al., 2003). Models that take into account the spatial distribution of diseases can be formulated by using the Poisson distribution to describe the number of cases in certain areas. In this situation, each area of investigation has its own risk factor and the Poisson model can be applied considering or not some covariates. Some references related to this subject are Lima (2004), Wartenberg (2001), Mollié (1999) and Ripley (1991). Another approach is the analysis of the data by means of geostatistics (Chilès and Delfiner, 1999; Griffith and Layne, 1999; Diggle and Ribeiro Jr., 2000).

In this paper, we will present a geostatistical model using the number of AIDS cases observed in the 1996 to 1999 period, in the cities of Minas Gerais state, in Brazil. The main objective is to present a statistical model that can be used as an alternative to obtain point and interval estimates for the true expected number of new cases of AIDS for cities which do not belong to the initial sample or for cities that did not present any diagnosed cases in the 1996 to 1999 period. A comparison will be performed with the geostatistical model proposed in Mingoti and Pantuzzo (1998) which was derived by using the number of diagnosed cases of AIDS observed in the 1990 to 1995 period, in the cities of Minas Gerais state.

2 Data presentation

The state of Minas Gerais is located in the southeastern area of Brazil which has about 17,295,955 inhabitants, according to the 1999 census. It possesses 853 municipal districts of which 339 had presented at least one diagnosed case of AIDS in the 1996-1999 period while the remaining 514 did not present any diagnosed cases in the same period. These cities will be herein referred to as “null cases”. Figure 1 shows the spatial distribution of the municipal districts investigated in this paper. The map shows the incidence of AIDS per 100,000 inhabitants. The cities are more concentrated in the central and southern areas of Minas Gerais state. The spatial distribution of the municipal districts that presented at least one case of AIDS diagnosed in the 1996-1999 period and the cities with "null cases" are also presented in Figure 1. Most of the cities with diagnosed cases

are located in the south and southwest of the state, which are the regions with a larger population.

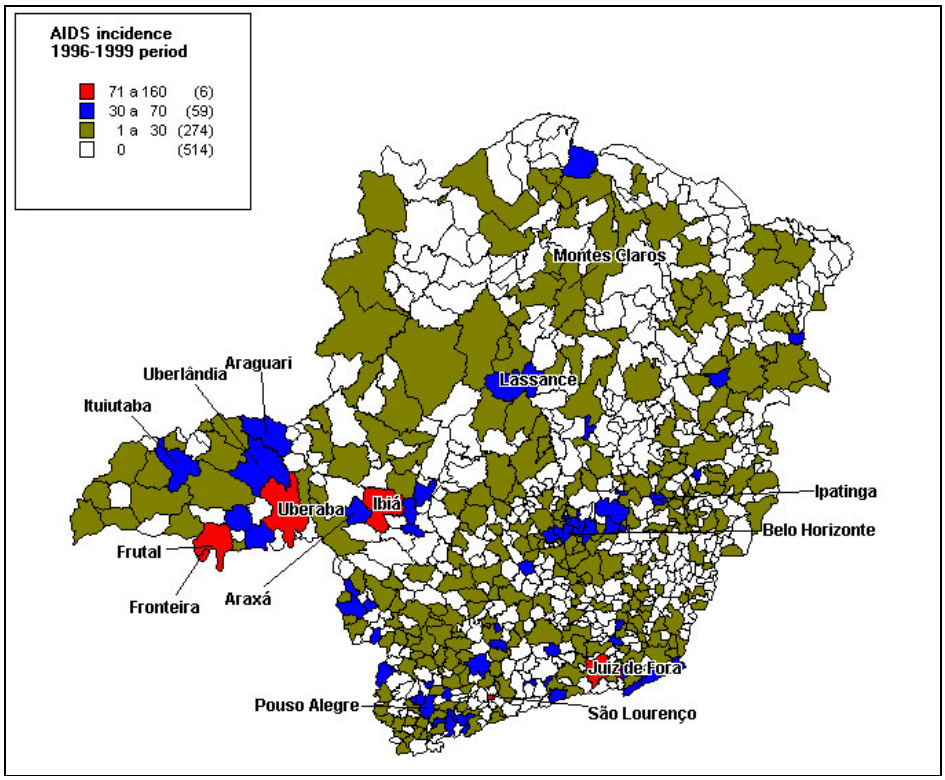

Figure 1 - Spatial distribution of AIDS incidence (number of cases per 100,000 inhabitants) of the municipal districts of Minas Gerais state in the 1996-1999 period.

Due to the variation among the municipal districts in terms of population size, the total number of diagnosed AIDS cases was divided by the respective population of each city. This variable will be referred to as "rate of cases". As an illustration, Figure 2 shows variable "rate of cases" as function of the latitude and longitude coordinates values of the cities. It can be observed that the regions with the highest rates are in the mid-south, located between the -19.9 and -20.0 latitude and between -41.0 and -45.0 longitude; and the Mineiro-Triangle (*Triângulo Mineiro*) located between -18.0 and -20.0 latitude and -46.0 and -49.0 longitude. The main cities in these two regions are *Belo Horizonte*, which is the capital of the Minas Gerais state and is the city with the highest population, *Juiz de Fora* and *Poços de Caldas* in the mid-south, and Uberaba and Uberlândia in the Mineiro-Triangle. The lowest rates occurred in the northeast of the state, between -14.0 and -15.0 latitude and -44.0 and -47.0 longitude. The city of *Bom Despacho* is located in this region and it is the one with the lowest rate.

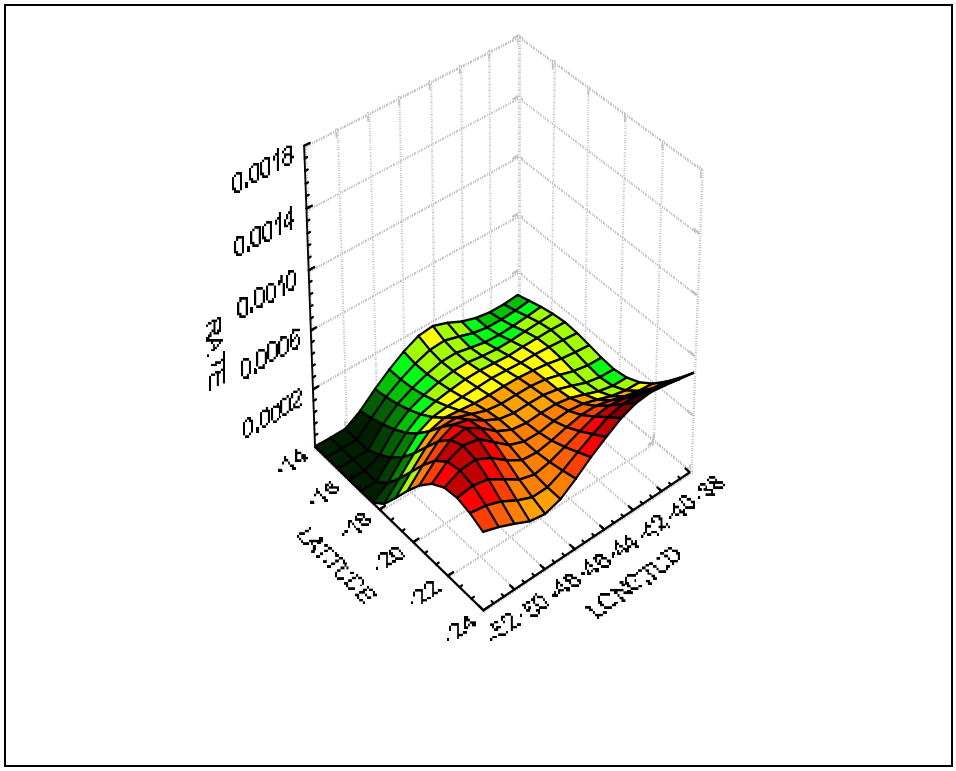

Figure 2 - Rate of AIDS cases as a function of latitude and longitude coordinates.

3 The geostatistics methodology

The use of geostatistics methodology is very common in geological studies (Isaaks and Srivastava, 1989) but can also be used in other contexts to describe regional variables such as sudden infant deaths in North Carolina (Cressie, 1993) or pediatric lead poisoning in Syracuse, New York (Griffith and Layne, 1999), for example. It is basically a set of statistical models used to describe the variation of quantitative random variables distributed in space or time or both. Therefore, every sampling unit in the data set is indexed by some coordinates such as the localization in space, in time or both. The main purpose is to use a statistical model to predict the value of the random variable under investigation for coordinates (or localizations) that do not belong to the sample data. Briefly speaking, the statistical analysis involves an identification of the semi-variogram model that best describes the variation of the random variable of interest among different localizations in the domain of study; the estimation of the parameters of the model; the validation of the chosen model, and the prediction for localizations which are not in the sample. Geostatistics is very useful in situations where the variables of interest are complex, difficult or very expensive to measure. Next, we will present the geostatistical analysis that was performed for the AIDS data set.

3.1 The data set and assumptions

For each one of the cities with diagnosed cases in the 1996-1999 period (Figure 1), let $s_j = (lat_j, long_j)$ be the vector with the respective latitude and longitude coordinates of the city j, j = 1, 2,... , 339. For each vector s_j , let $Y(.)$ be the logarithm of the variable "rate of cases" defined in section 2.0. The distribution of the transformed variable $Y(.)$ is approximately normal, which allows for the use of the geostatistics methodology (see Figures 3a. and 3b.). For the analysis presented in this paper, only the cities with at least one case were considered because the purpose of the paper is to predict the number of cases for cities that did not have any reported cases in the 1996-1999 period. In any case, since the logarithm transformation is not possible for cities with zero cases, any analysis using these cities would require some correction.

The main objective is to describe the spatial variability of $Y(.)$ using the 339 observations (cities) in the sample. For each pair of locations $s_l \neq s_k$, $Y(s_l)$ and $Y(s_k)$ are expected to be correlated in such way that the correlation decreases as s_l and s_k are far apart. Let D be the investigation area (here D=state of Minas Gerais), then $\{Y(s), s \in D\}$ is a stochastic process (Cressie,1993). In the analysis by means of Geostatistics it is necessary to impose that the stochastic process $\{Y(s), s \in D\}$ be intrinsically stationary and isotropic, i.e.,

i) $E[Y(s)] = \mu$, $\forall\, s \in D$

ii) $\mathrm{Var}[Y(s_l) - Y(s_k)] = 2\gamma(h)$ where $s_l \neq s_k \in$ D, and $\left\| s_l - s_k \right\| = h$.

Basically, these assumptions indicate that $Y(.)$ has a constant average in D, and for each pair $s_l \neq s_k \in$ D, the variance of the differences $[Y(s_l) - Y(s_k)]$ is just a function of the distance between the s_l and s_k coordinates. Quantities $2\gamma(.)$ and $\gamma(.)$ are called, respectively, variogram and semi-variogram of the process $\{Y(s), s \in D\}$. Some of the more common models of theoretical variograms are the linear, spherical, exponential and gaussian variograms. For more details see Cressie,1993.

To perfom the analysis of any possible trends, Figures 4 and 5 present the variable lograte $Y(.)$ plotted against latitude (north-south) and longitude (east-west). The city of *Belo Horizonte* is hightlighted because is an outlier. Other plots (not shown in this paper) were made without Belo Horizonte city and there was no visible trend in either of the two directions. Therefore, the AIDS incidence process $\{Y(s), s \in D\}$ was considered stationary in mean. The isotropy condition will be discussed in the next section.

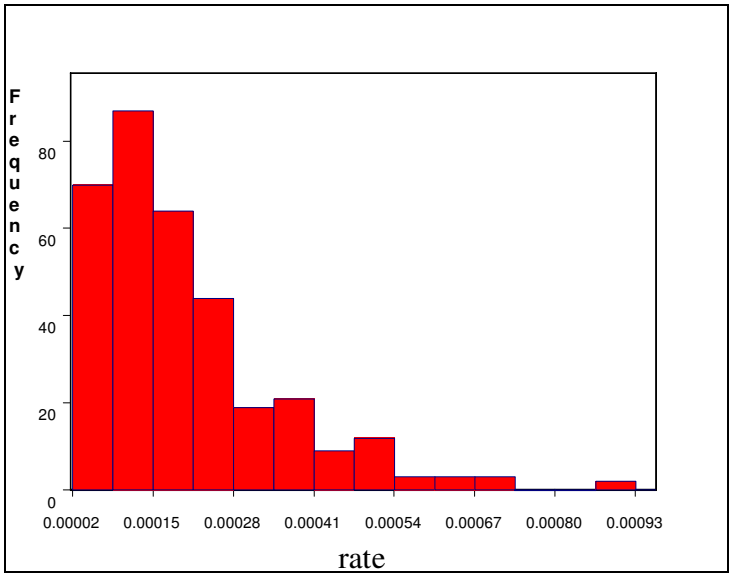

Figure 3a - Distribution of rate of cases.

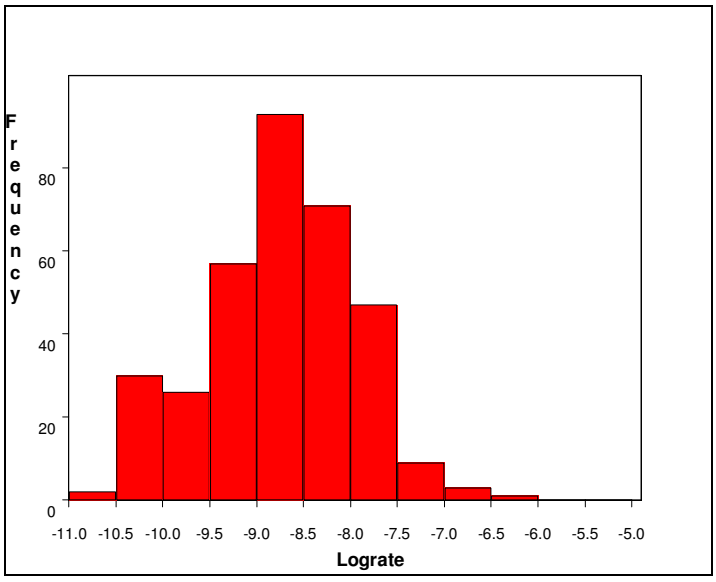

Figure 3b - Distribution of the logarithm of rate of cases.

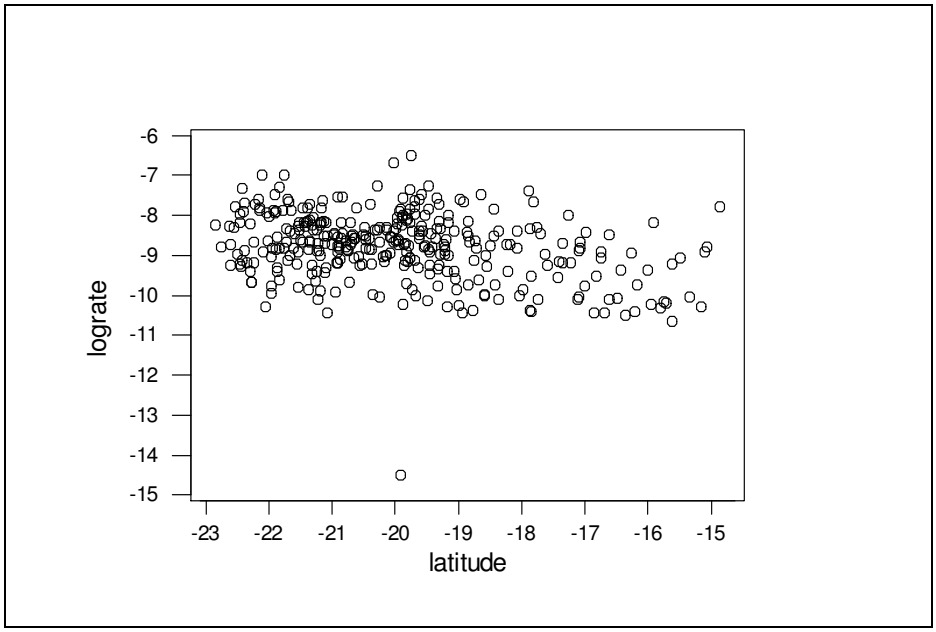

Figure 4 - Variable lograte $Y(.)$ versus latitude coordinate.

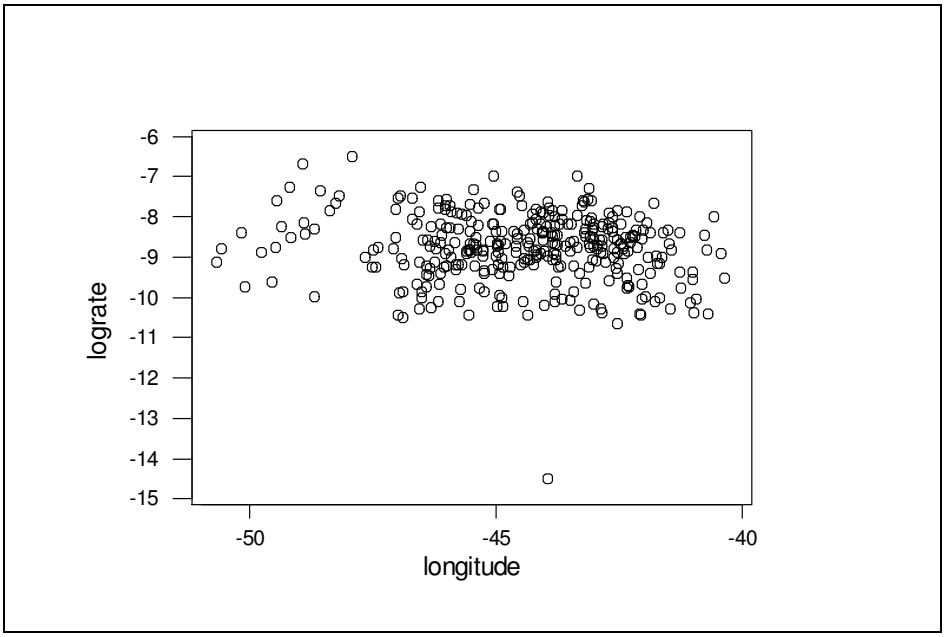

Figure 5 - Variable lograte $Y(.)$ versus longitude coordinate.

3.2 Variogram estimation

The theoretical variogram of the AIDS incidence process $\{Y(s), s \in D\}$ was estimated by using the classical estimator proposed by Matheron (1963) defined as:

$$2\hat{\gamma}(h) = \frac{\sum_{N(h)} \left[Y(s_l) - Y(s_k) \right]^2}{\left| N(h) \right|} \quad , \quad h \in D \tag{3.2.1}$$

where $N(h) = \{ (s_l, s_k): \| s_l - s_k \| = h;\ l, k = 1,\ldots,n,\ l \neq k \}$, and $|N(h)|$ is the cardinality of the $N(h)$ set. Other estimators can be used such as the robust, the median and the nonparametric (Cressie (1993), or those proposed by Genton (1998) and Delay and Marsily (1994) for example. The variogram given by equation (3.2.1) is called the experimental variogram.

Figure 6 presents the experimental variogram of $Y(.)$ in the north-south direction and the fitted spherical model. When plotted in the east-west, northeast-Southwest and northwest-southeast directions, the graphic form of the experimental variogram was very similar to the variogram presented in Figure 6, which was an indication that the lograte process $\{Y(s), s \in D\}$ was approximately isotropic.

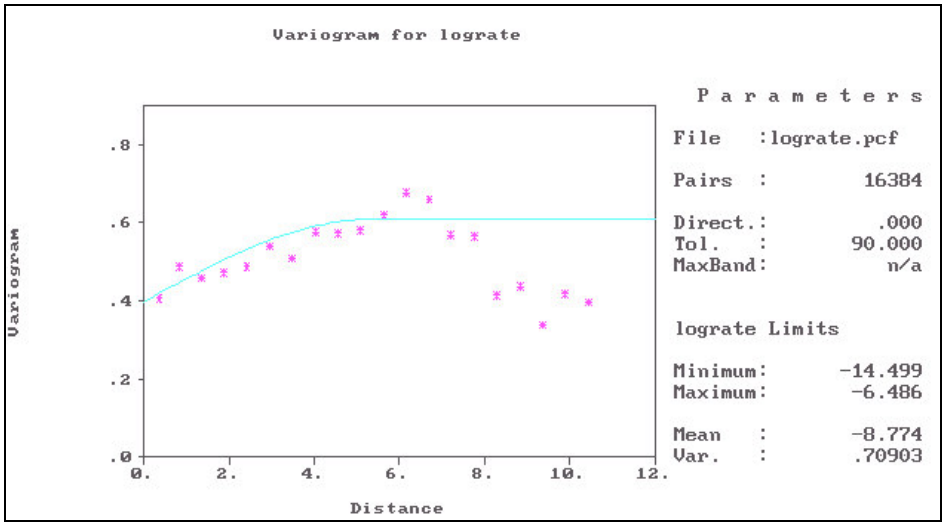

Figure 6 - Experimental variogram of the lograte $Y(.)$ variable.

The parameters of the spherical model were estimated by weighed minimum squares (Cressie, 1985) and the fitted equation is given in (3.2.2). Constant $\hat{c}_0 = 0.398$ is called nugget effect (Matheron, 1963) and is the point where the curve touches the vertical axis; constant $\hat{c}_s = 0.211$ is the partial sill and $\hat{a}_s = 5.343$ is the range of the model. The total

sill is constant $\hat{c}_0 + \hat{c}_s$ = 0.609 and it is an estimate of variance σ^2 of the {*Y*(*s*), *s* ∈ D} AIDS process. For cities thet are at least 5.343 away, there is no more correlation and the only effect affecting response *Y*(.) is natural variance σ^2.

The last eight points of the experimental variogram were not considered in the estimation of parameters of the spherical model because they were based on a very small number of city pairs.

$$\gamma(h;\theta) = \begin{cases} 0 \quad if \ \| h \| = 0 \\ 0.398 + 0.211 \left\{ \frac{3}{2} \left(\frac{\| h \|}{5.343} \right) - \frac{1}{2} \left(\frac{\| h \|}{5.343} \right)^3 \right\} \quad if \quad 0 < \| h \| \leq 5.343 \\ 0.609 \quad if \quad \| h \| \geq 5.343 \end{cases} \qquad (3.2.2)$$

3.3 Predicting new locations - ordinary kriging

Ordinary Kriging is a statistical method used to forecast values of the intrinsically stationary and isotropic Gaussian stochastic process for locations or areas which do not belong to the observed sample (Cressie, 1990). The prediction for a new location s_0 is a weighed average of the sampled observations from a neighborhood of s_0. Briefly speaking, if $Y(s_1), Y(s_2), ..., Y(s_n)$ are the sample values of the variable lograte *Y*(.) for cities of Minas Gerais state located at $s_1, s_2, ..., s_n$, and if s_0 is a city in the state which was not previously sampled then, unbiased linear predictor $\hat{Y}(s_0)$ for true value $Y(s_0)$ which minimizes the prediction mean error square, $E[(Y(s_0) - \hat{Y}(s_0)]^2$, is given by the equation:

$$\hat{Y}(s_0) = \sum_{i=1}^{n} \lambda_i Y(s_i) \quad , \quad s_0 \in D \qquad (3.3.1)$$

where $\sum_{i=1}^{n} \lambda_i = 1$, and weights λ_i are obtained by solving the equations:

$$\lambda_0 = \Gamma_0^{-1} \gamma_0 \qquad (3.3.2)$$

where

$$\lambda_0 \equiv (\lambda_1, \lambda_2, ..., \lambda_n, m)' \ ,$$

$$\lambda_0 \ \gamma_0 \equiv \lambda_0 \ (\gamma(s_0 - s_1), \gamma(s_0 - s_2), \ ..., \gamma(s_0 - s_n), 1)'$$

Γ_0^{-1} is the inverse of matrix (*n+1*)*x*(*n+1*) matrix Γ_0 defined as:

$$\Gamma_0 = \begin{cases} \gamma(s_i - s_j), & i = 1,...,n;\ j = 1,...,n \\ 1 \quad , & i = n+1;\ j = 1,...,n \\ 0 \quad , & i = n+1;\ j = n+1 \\ 1 \quad , & i = 1,...,n;\ j = n+1 \end{cases} \tag{3.3.3}$$

Parameter m is a Lagrange multiplier used in the minimization procedure so that weights λ_i will satisfy the condition $\sum_{i=1}^{n} \lambda_i = 1$. The variance of the prediction error is given by $\sigma_e^2(s_0) = \lambda_0' \gamma_0$, and an interval with 95% of confidence for the true value $Y(s_0)$ is defined as: $(\hat{Y}(s_0) \pm 1{,}96\, \sigma_e(s_0))$.

In practice the forecast of value $Y(s_0)$ is performed using all the sample points which belong to an ellipse centered at location s_0. Therefore, the user can control the neighborhood to be used in the prediction of $Y(s_0)$ by choosing the axes, the radius and the orientation of the ellipse.

The validation of the prediction model is also performed using equations (3.3.1) to (3.3.3). Each one of the 339 sampled locations was taken out of the data set at a time. For each specific location the forecast was performed and the standardized prediction error calculated (Cressie, 1993). The distribution of the prediction errors (residuals) was approximately Normal with mean 0.031 and standard deviation equal to 1.167. The analysis showed no spatial pattern of the residuals. Therefore, under this aspect the proposed prediction model was considered adequate to describe AIDS lograte variable $Y(.)$.

3.4 Prediction of the number of cases

Although the Geostatistical model was adjusted for lograte variable $Y(.)$, our main objetive was to predict the true number of diagnosed cases of AIDS for each municipal district (let us called it a $Z(.)$ variable). To perform the forecast of $Z(.)$, the inverse transformation with the lognormal correction (Mood, 1974) was used. For every location with coordinates s_0, the prediction value for $Z(.)$ was generated by using the equation:

$$\hat{Z}(s_0) = \hat{X}(s_0)\, x\, pop(s_0)$$

where

$$\hat{X}(s_0) = \exp\{\hat{Y}(s_0) + \frac{1}{2}\sigma_e^2(s_0)\}$$

and $pop(s_0)$ represents the total population of the municipal district located at coordinates s_0. The 95% confidence interval for the true value of $Z(s_0)$ is given by:

$$\left\{ \begin{array}{l} pop\ (s_0)\ \exp\left[\hat{Y}(s_0) - 1{,}96\,\sigma_e(s_0)\right]; \\ pop\ (s_0)\ \exp\left[\hat{Y}(s_0) + 1{,}96\,\sigma_e(s_0)\right] \end{array} \right\}$$

As an illustration, we present the prediction for the number of diagnosed cases of AIDS for some cities of Minas Gerais in Table 1. The cities that were considered as outliers are naturally more difficult to predict. Therefore, larger prediction errors are expected for *Belo Horizonte, Juiz de Fora, Uberaba* and *Uberlândia*.

Table 1 - Prediction for the total number of diagnosed cases of AIDS for some cities of Minas Gerais state, Brazil, using the geostatistical model

Cities	True observed value	Prediction value	Confidence interval (95%)
1. Barbacena	18	25.65	(5 ; 77)
2. Belo Horizonte (*)	1056	795.69	(169 ; 2380)
3. Betim	80	62.97	(13 ; 188)
4. Contagem	183	75.59	(16 ; 226)
5. Curvelo	7	10.76	(2 ; 33)
6. Diamantina	7	6.63	(1 ; 20)
7. Divinópolis	40	25.67	(5 ; 78)
8. Governador Valadares	68	54.13	(11 ; 164)
9. Ipatinga	77	49.18	(10 ; 147)
10. Itajubá	53	22.03	(5 ; 66)
11. Juiz de Fora	392	129.36	(27 ; 390)
12. Lavras	17	22.70	(5 ; 68)
13. Montes Claros	32	28.23	(6 ; 86)
14. Paracatu	7	5.25	(1 ; 17)
15. Poços de Caldas	46	16.33	(3 ; 50)
16. São João Del-Rei	22	14.64	(3 ; 44)
17. Teófilo Otôni	33	22.40	(5 ; 68)
18. Uberaba (*)	362	75.90	(15 ; 232)
19. Uberlândia (*)	211	109.33	(22 ; 334)

(*) These cities were considered as outliers.

The prediction values for some cities which did not present any diagnosed cases in the 1996-1999 period are shown in Table 2 ("null cases"). They can be viewed as estimates of the total number of diagnosed cases that would be expected to occur in the next 4 years for those cities with "null cases" in 1996-1999. Considering the prediction performed for all the 514 "null cases" municipal districts of Minas Gerais, the number of new diagnosed cases expected for those cities in the 2000-2003 period was 784. If we take into account that the mechanisms used by the health institutions to report AIDS cases to government offices are not perfect, that there is a delay for the information of a reported case to be available in the official database and that many AIDS cases are not reported at

all due to social reasons, the predictions for the "null cases" cities could also be used as a roughly estimate measure of underreporting of AIDS cases in the Minas Gerais state. It is important to point out this is only a rough estimate since better statistical methods to estimate underreporting cases are available in the literature (Caiaffa, Mingoti et. al. 2003)

Table 2 - Prediction for the total number of diagnosed AIDS cases for some cities of Minas Gerais which did not present any diagnosed cases in the 1996 -1999 period ("null cases")

Cities	Population	Prediction	Prediction (rounded)	Confidence Interval (95%)
1. Arinos	17149	1.2139	2	(0; 3)
2. Brasília de Minas	43615	3.6080	4	(2; 6)
3. Espinosa	31550	2.4147	3	(1; 4)
4. Minas Novas	33446	2.5328	3	(1; 4)
5. Lagoa Formosa	15901	1.8097	2	(1; 3)
6. Abaeté	22082	2.6420	3	(1; 4)
7. Paraopeba	18623	2.8550	3	(1; 5)
8. Guanhães	26518	4.2401	5	(2; 7)
9. Inhapim	33288	5.7974	6	(4; 8)
10. Bambuí	21187	4.8447	5	(3; 7)
11. Cláudio	20530	3.2047	4	(2; 5)
12. Ouro Fino	28320	4.5987	5	(3; 7)
13. Nepomuceno	24771	4.3545	5	(3; 6)
14. Tiradentes	11695	2.0481	3	(1; 3)
15. Raul Soares	29036	3.9011	4	(2; 6)

4 Looking back to the predictions made for the 1996-1999 period using the observed AIDS cases diagnosed in the 1990-1995 period

Mingoti and Pantuzzo (1998) presented a geostatistical model to predict the number of diagnosed cases of AIDS in Minas Gerais state for the 1996-2001 period. In that paper, the semi-variogram model used to generate the predicions for the $Y(.)$ lograte variable was a spherical model with parameters: nugget=0.35, partial sill=0.47, range=8.50. Estimated variance σ^2 was 0.82. Their analysis was based in a sample with 293 municipal districts that presented diagnosed cases of AIDS in the 1990-1995 period. As it can be seen, the parameters of the spherical model changed from 1990-1995 to 1996-1999 and variance σ^2 decreased. The distance for which the correlation between the cities is nearly zero decreased in the 1996-1999 period.

The experimental variogram was also calculated considering the 1996-1999 data only for the cities that had presented diagnosed cases of AIDS in the 1990-1995 period (i.e. 293 cities). The estimated variogram for the lograte variable $Y(.)$ was a spherical model with nugget=0.31, partial sill=0.43, range=8.24 and estimated variance σ^2 equal to 0.74. Therefore, the estimated parameters were very similar to those of the spherical semi-

variogram model fitted by using the data from 1990-1995. This is an indication that the spatial behaviour of the $Y(.)$ lograte variable did not change much in both periods of investigation for those 293 municipal districts. However, the introduction of new districts in the sample had the effect of decreasing the variance of the AIDS $Y(.)$ lograte and the distance for which the correlation between cities is negligible.

A comparison between the projections obtained in Mingoti and Pantuzzo's paper and the true number of diagnosed AIDS cases observed in the 1996-1999 period was performed. Considering the projections for the set containing the 293 cities with at least one diagnosed AIDS case in the 1990-1995 period, the prediction mean error (*ME*) was equal to -8.23 and a prediction square root of a square error mean (*SRSEM*) was equal to 52.61. About 88% of the 95% confidence intervals had covered the true number of diagnosed AIDS cases observed in the respctive city in the 1996-1999 period. When the cities considered as outliers, *Belo Horizonte*, *Juiz de Fora*, *Uberaba* e *Uberlândia*, were omitted from the sample, *ME* and *SRSME* were equal to -3.02 and 10.76, respectively.

Considering the projections for the set with 463 cities which did not present any cases in the 1990-1995 period ("null cases") an *ME* equal to 1.24 and an *SRSME* equal to 1.84 were found. About 91% of of the 95% confidence intervals covered the true number of diagnosed AIDS cases observed in the "null cases" cities in the 1996-1999 period.

Therefore, the analysis of prediction errors presented here, i.e. the validation of the geostatistical projection model, gives an indication that the geostatistical model constructed in Mingoti and Pantuzzo's paper for the AIDS diagnosed cases of the Minas Gerais state showed a good performance.

5 Final comments

At the beginning of this paper, it was pointed out that the main objective was to investigate the geostatistics methodology to find out if it would give good predictions for the number of diagnosed AIDS cases in Minas Gerais state. It is true that the presented geostatistical model does not take into consideration important variables such as the incubation period of the disease, different mechanisms of transmission and infection, social class, among other covariates. Also, it is well known that not all the diagnosed AIDS cases are reported by the Brazilian Health institutions. However, the results presented in this paper show that the geostatistical model is a good alternative prediction model with an advantage of being very simple to apply. The comparison of the predictions presented in Mingoti and Pantuzzo's paper (1998) obtained by using data from 1990-1995 with the true values for the 1996-1999 period provides a good support for this assertion. Clearly, the proposed model can be improved by adding covariates or by using the Cokriging technique (Cressie, 1993; Chilès and Delfiner, 1999). However, its simplicity makes it very appealing.

There are some cities that naturally are more difficult to predict. For the outliers, an alternative is to estimate the number of cases as the upper limit of the 95% confidence interval or to construct other statistical models.

A proposal to improve the prediction models is to consider time as a location variable (Kyriakidis and Journel, 1999). For the data set that we analysed this alternative was not possible because many of the cities did not have diagnosed cases every year.

Therefore, the frequency of "zero" values was too high for a time variation study in the geostatistics context.

It is important to point out that an immediate application of the developed geostatistical model refers to the prediction of the number of cases for the cities that did not present any cases in the 1996-1999 period ("null cases"). The predictions for these cities can be viewed as the number of new cases that will be diagnosed in the 2000-2003 period.

Finally, the analysis presented in this paper reeinforces the fact that the information about the spatial location of the cities in the sample is of extreme relevance for the construction of statistical models to project the occurrence of AIDS cases.

MINGOTI, S. A.; LEITE, A., G.; ROSA, G. Descrevendo o número total de casos diagnosticados de AIDS por meio de geoestatística. *Rev. Mat. Estat.*, São Paulo, v.24, n.1, p.61-76, 2006.

- *RESUMO: Neste artigo um modelo geoestatístico é ajustado para o número total de casos diagnosticados de AIDS nos municípios de Minas Gerais no período de 1996 a 1999. O modelo de previsão é usado para estimar o número de casos diagnosticados e não notificados aos órgãos de vigilância da Saúde Pública de Minas Gerais. É mostrado que não há muita diferença entre o modelo de semi-variograma apresentado neste artigo e aquele encontrado em Mingoti e Pantuzzo (1998) e que foi construido com dados de número de casos diagnosticados de AIDS nos municípios de Minas Gerais no período de 1990 a 1995. Os resultados que são apresentados neste artigo indicam que o modelo geoestatístico tem um bom desempenho na descrição do número total de casos diagnosticados de AIDS em Minas Gerais.*

- *PALAVRAS-CHAVE: Geoestatistica; semi-variograma; predição; AIDS; Minas Gerais.*

References

BECKER, N. G.; MARSCHNER, I. C. A method for estimating age specific relative risk of HIV infection from AIDS incidence data. *Biometrika*, London, v.80, p.165-178, 1993.

BROOKMEYER, R. AIDS, epidemics, and statistics. *Biometrics*, Washington, v.52, p.781-796, 1996.

BROOKMEYER, R.; GAIL, M. H. A method for obtaining short-term projections and lower bounds on the size of the Aids epidemic. *J. Am. Stat. Assoc.*, New York, v.83, p.301-308, 1988.

BROOKMEYER, R. , LIAO, J. Statistical modelling of the AIDS epidemic for forecasting health care needs. *Biometrics,* Washington, v.46, p.1151-1163, 1990.

CAIAFFA, W. T. et. al. Estimation of the number of injecting drug users (IDUs) attending an outreach syringe exchange program (SEP), and the infection with human immunodeficiency virus (HIV) and hepatitis C virus (HCV): the AjUDE-BRASIL project, *J. Urban Health*, Cary, v.80, p.106-114,2003.

CHILÈS, J, P. ; DELFINER, P. Geostatistics. New York: John Wiley, 1999. 695p.

CRESSIE, N. *Statistics for spatial data*. New York: Wiley, 1993, 900p.

CRESSIE, N. The origins of Kriging. *Math. Geol.,* New York, v.22, n.3, p.239-252, 1990.

CRESSIE, N. Fitting variogram models by weighted least squares. *J. Int. Assoc. Math. Geol.*, New York, v.17, n.5, p.563-586, 1995.

DELAY, F.; MARSILY, G. The integral of the semivariogram: a powerful method for adjusting the semivariogram in geostatistics. *Math.l Geol.*, New York, v.26, n.3, p.301-321, 1994.

DIGGLE, P. J.; RIBEIRO Jr., P. J. *Model based geostatistics*. In: SINAPE, 14., 2000, São Paulo. *Resumos*...São Paulo: ABE. 2000. 129p.

DING, A. A.; WU, H. A comparison study of models and fitting procedures for biphasic viral dynamics in HIV-1 infected patients treated with antiviral therapies.*Biometrics*, Washington, v.56, p.293-300, 2000.

FAUCETT, C. L.; SCHENKER, N.; TAYLOR, J. M. G. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics*, Washington, v.58, p.37-47, 2002.

GENTON, M. G. Highly robust variogram estimation. *Math. Geol.*, New York, v.30, n.2, p.213-221, 1998.

GRIFFITH, D. A.; LAYNE, L. J. *A casebook for spatial statistical data analysis*. New York: Oxford University Press, 1999. 506p.

ISAAKS, E. H.; SRIVASTAVA, R. M. *Applied geostatistics*. New York: Oxford University Press, 1989, 561p.

JAGER, J. C.; RUITENBERG, E. J. *Statistical analysis and mathematical modelling of AIDS*. New York: Oxford University Press, 1988, 167p.

LIMA, M. S. *Avaliação do poder do teste da estatística scan para múltiplos clusters*. 2004. 84f. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

KYRIAKIDIS, P. C.; JOURNEL, A. G. Geostatistical space- time models: a review. *Math. Geol.*, New York, v.31, n.6, p.651-684 , 1999.

MATHERON, G., Principles of geostatistics. *Econ. Geol*., Lancaster, v.58, p.1246 -1266, 1963.

MCEVOY, M.; TILLET, H. E. Some problems in the prediction of future numbers of cases of the acquired immunodeficiency syndrome in the U.K. *Lancet*, Boston, v.2, p.541-542, 1985.

MAYER-HAMBLETT, N.; SELF, S. A regression modelling approach for describing patterns of HIV genetic variation. *Biometrics*, Washington, v.57, p.449-460, 2001.

MOLLIÉ, A. Bayesian and empirical bayes approaches to disease mapping. In: DISEASE MAPPING AND RISK ASSESSMENT FOR PUBLIC HEALTH. New York: John Wiley, 1999. p.120-142.

MINGOTI, S. A.; PANTUZZO, A. E. Predição do número total de casos diagnosticados de AIDS dos municípios de Minas Gerais através de técnicas de estatística espacial. *Rev. Mat. Est.*, São Paulo, v.16, p.59-79, 1998.

MOOD, A.M.; GRAYBILL, F.; BOES, D. Introduction to theory of statistics. New York: McGraw- Hill, 1974. 564p.

MORGAN, W. M. ; CURRAN, J. W. Acquired immunodeficiency syndrome: current and future trends. *Public Health Rep*., Washington, v. 101, p. 459-464, 1986.

ROSENBERG, P. The scope of Aids epidemic in the United States. *Science*, Washington, v.270, p.1372-1375, 1995.

RIPLEY, B. D. Statistical inference for spatial processes. Cambridge: Cambridge University Press, 1991. 148p.

SOLOMON, P. J., WILSON, S. R. Accommodating change due to treatment in the method of back projection for estimating HIV infection incidence. *Biometrics*, Washington, v.46, p.1165-1170, 1990.

WARTENBERG, D. Investigating disease clusters: why, when and how? *J. R. Stat. Soc.* Ser. A, London, v.164, p.13-22, 2001.

WU, H.; DING, A. A. Population HIV-1 dynamics in vivo: applicable models and inferential tools for viorological data from aids clinical trials. *Biometrics,* Washington, v.55, p.410-418, 1999.

Recebido em 15.06.2004.

Aprovado após revisão em 10.02.2006.