

Spatial interpolation vs neural network propagation as a method of extrapolating from field surveys

Weir-Smith, G. and Schwabe, C.A.

GIS Centre, HSRC (Human Sciences Research Council), Pretoria

ABSTRACT

In a rapid changing society like South Africa, up to date and accurate information on the socio-economic, service delivery, demographic, substance abuse, disease and other conditions of the nation is needed on a regular basis. The gathering of representative information in a vast country like South Africa can be a costly exercise. A well-designed sample, however, can lay the base for the collection of information and for further research and analysis. Utilising a good sample design together with a spatial extrapolation method provides an alternative to frequent and expensive surveys. In order to save on the cost of surveys the HSRC (Human Sciences Research Council) GIS Centre is using extrapolation methods via neural networks to assign values – based on existing socio-economic and demographic information - to all entities not included in a sample survey. Another methodology that can be utilised in a GIS environment is spatial interpolation. This methodology estimate values for unsampled areas based on the values of surrounding sampled points. Both mentioned methods will save significant costs in terms of gathering of information. This paper aims to focus on the spatial interpolation method and compare results with neural network propagation outputs from a survey done on substance abuse behaviour linked to crime patterns in selected police stations throughout the country. This methodology provides comprehensive spatial information for decision makers in South Africa to make accurate and timely decisions about threatening social conditions like poverty, substance abuse, unemployment, crime and HIV/AIDS.

INTRODUCTION

Field surveys are costly exercises and the use of such data is often limited to a pre-defined geographic area. The spatial value of field survey results are regularly under utilised due to restricted application by users. The HSRC conducts several surveys a year and in the past couple of years artificial neural network propagation was used to extrapolate survey results to a more representative or national level. The aim of this paper is to examine the methodology of both models and to compare the results of spatial interpolation done on a recent substance abuse survey.

BACKGROUND DATA

The HSRC conducted a national survey on substance abuse patterns in 2000. A sample of 150 (out of 1089) police stations nationally were drawn using a stratified probability sample and arrestees at these stations were interviewed about their substance abuse and related crime behaviour. The results of these findings will be used by the two methodologies mentioned above.

NEURAL NETWORKS

Extrapolation refers to the estimation of values for unsampled points which lie outside the boundary of an existing sample set of data (AGI 1999). The extrapolation of sampled data is done to achieve a more complete analysis of a selected data set. In the past the HSRC GIS Centre used neural network propagation to obtain such results. The artificial neural networks are based on the structure and functioning of the human brain and consist of a large number of simple processing units known as neurons (Singh & Treleaven 1998: 2). Using the back propagation method the neural network software is able to use existing data to provide an output data set for non-sampled points ("Output" in Figure 1 below).

In the past the result was extremely successful and verified by experts in the field of the particular application. The process of neural network propagation is often time consuming, especially if big data sets are being used. The process can be explained by the illustration below.

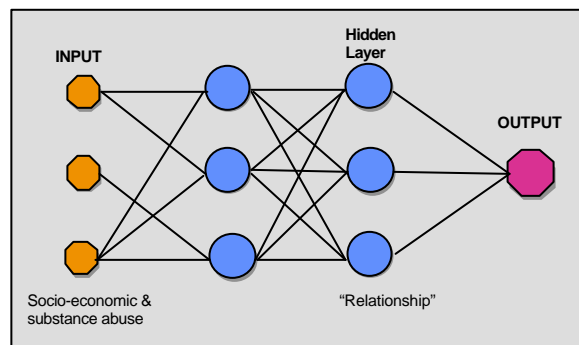


Figure 1

The neural network software uses input data from a sample to create estimated values for non-sampled areas. In the case of substance abuse research

this input data set consisted of a substance abuse data set for a specific area (e.g. a sample) as well as a socio-economic data set matching the area for which the estimated substance abuse figures is needed (the "input" on Figure 1). The latter area is usually of provincial or national coverage. The software develops intricate relationships ("hidden layer" in Figure 1) between the two data sets in order to understand the factors determining the use and abuse of substances (the outcome). In other words it uses the familiar (socio-economic conditions for the whole as well as substance abuse results from the sample) to determine what the likelihood of substance abuse (the unfamiliar) would be in non-sampled areas. Once this understanding is completed the software is able to provide an output data set for the desired area. The output data set consists of estimated substance abuse values for the "non-sampled" area whether it is a province or at a national level.

By using this method powerful data sets can be created using sampled data to extrapolate to a universe. The resulting data set provides an indication of what the expected outcome for specific variables will be.

Extrapolated results based on this method are shown in Figure 2 below. Percentages of substance use were created in the original data set. This was used together with socio-economic data to determine what the possible substance use would be for areas not sampled. (In this case police station boundaries were identified as the primary spatial building block.) For the purpose of this analysis data from the Gauteng province only will be used, although the survey was done nationally.

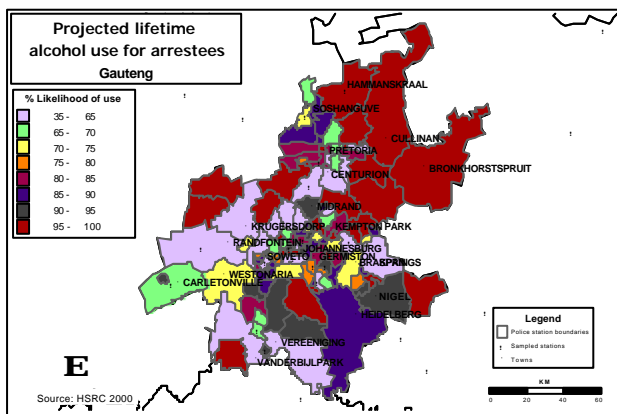


Figure 2

No statistical validation method exists for this method, because it is based on "artificial intelligence". Although neural networks can be used successfully for classification and pattern recognition it is not able to explain the reasoning used to arrive at a decision. In the past the HSRC has used experts in specific local areas and

familiar with the application field used, to verify the results. Based on success in the past this methodology has been accepted to be a true reflection of reality.

In order to explore new techniques that might result in time and cost savings as well as increase accuracy of data, spatial interpolation was considered. This methodology and results will be discussed in the next section.

SPATIAL INTERPOLATION

Interpolation refers to the estimation of Z values of a surface at an unsampled point based on the known Z values of surrounding points. (AGI 1999.) It is usually presented by isolines and mainly associated with physical features like rainfall, contours and temperature. This study is an attempt to apply the use of spatial interpolation to the human sciences. The specific methodology used refers to exact interpolators. (University of British Columbia, 1997).

The original survey results were used as extrapolation base. The data was imported into GIS software as containing X, Y and Z values. The Z value represented the specific field from the database which had to be interpolated. In this case lifetime alcohol use by arrestees was used as the Z value.

In order to create contours the data had to be triangulated first to form a surface model. This allowed the calculation of contours based on a planar surface created by the TIN (Triangulated Irregular Network) object. It is calculated by drawing non-overlapping, connecting triangles between all data points. Figure 3 illustrates the TIN object that was created for South Africa, but focussing on Gauteng only. The software used was TNT Mips.

The line breaks (contours) created on the edges of the triangles is clear from this figure.

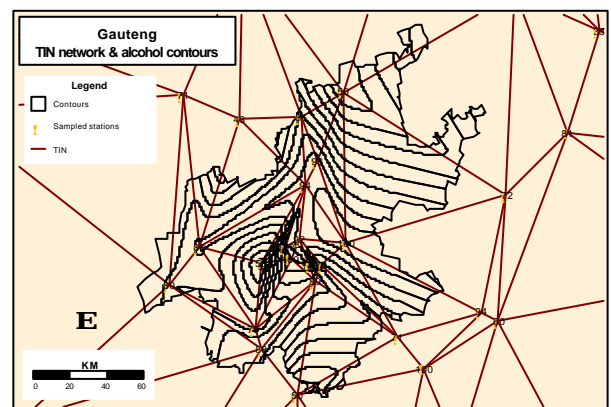


Figure 3

This TIN model was used to create contours at 5% intervals. The software offers two options of contouring from a TIN – linear or cubic. The linear method was selected and it calculated the contours by treating each TIN triangle as a planar surface. Contours are drawn across the sides of the connected, tilting triangular surfaces.

When a contour passes between two TIN nodes, the location of its intersection with the triangle edge is determined by linear interpolation from the node Z values. The closer the input points, the denser the contours would be. The software created vector line and polygon contours at the selected intervals. The linear method was selected because of its accuracy and precision as well as time saving.

VALUE OF INTERPOLATION

Contouring allows simple visual interpretation of data. The value of unsampled points can be estimated with ease. The methodology does however not realistically represent the natural surface or processes that might influence the natural surface. Another drawback is that it does not account for inaccuracies in the input data.

It is known that data measurement accuracy, data density, data distribution and spatial variability has the greatest influence on interpolation accuracy. (MacEachren and Davidson 1987: 312).

The visual impact of spatial interpolation based on social phenomena allows quick interpretation by readers and analysts alike. Underlying factors will however only be revealed by in depth analysis.

The two methods used to extrapolate from the survey data will be compared in the following sections. Firstly a visual comparison will be done and secondly it will be compared statistically.

VISUAL COMPARISON

Although it might not be a fair reflection in terms of data distribution, the same intervals were used for the maps below for the purposes of comparison.

Neural Networks

The neural network analysis ensures a value for individual spatial entities of the non-sampled points. (See Figure 2.) It is therefore accurate in providing statistics at a detailed level. This level of accurate detail depends of course on the representivity of the original sample.

Spatial interpretation of the Gauteng map in Figure 2 indicates that police stations of extremely high (95-100%) expected lifetime use of alcohol among arrestees are concentrated in the following police station precincts:

- ◆ East Rand – Devon, Dunnotar, Reigerpark, Sebenza and Kempton Park.
- ◆ North East – from Olifantsfontein to Cullinan (including Wonderboompoort in Pretoria).
- ◆ West – Erasmia, Muldersdrift, Fairland, Hekpoort and Magaliesburg.
- ◆ Central – Langlaagte and Johannesburg Central.
- ◆ South – Kliprivier, Edenpark and The Barrage.

If the second highest category (90-95%) of use is taken into consideration, the data demonstrates a strong likelihood of high alcohol use in the east, south and northeast of Gauteng.

The value of this data is that it can be interpreted in terms of the additional socio-economic variables that were used to calculate the estimated value. In other words the results not only refer to potential substance abuse, but also to other variables linked to the same spatial entity.

Spatial Interpolation

Results of the spatial interpolation is shown in Figure 4 below. The linear methodology used in TNT Mips produces angular contours. In order to obtain smoother effects the contours were splined using the Quadratic Bspline method in TNTMips. As a result of this step spatial accuracy might have been traded off for less angular contours.

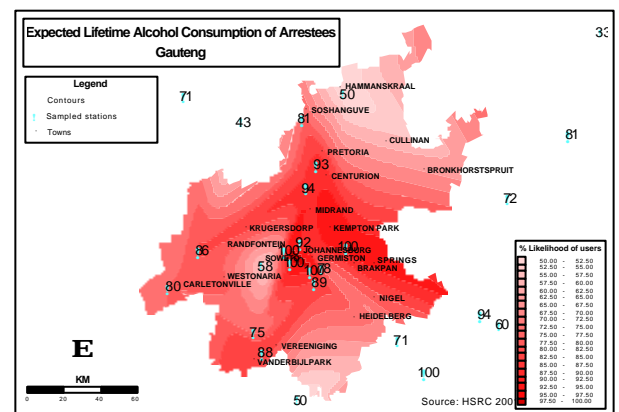


Figure 4

According to Figure 4 areas of high likelihood of alcohol consumption is situated in a band stretching from Springs in the south east through Kempton Park to Midrand as well as to the south of Johannesburg central.

Contrary to the neural network result the north east of the province is not identified as an extremely high likelihood of use area. In fact the area ranges from a 75 to 50% likelihood of lifetime use. The same applies to the West Rand which was allocated a lower likelihood of use, namely 75% (in comparison to the 95%+ of the neural network result.)

A secondary band of high likelihood (85-90%) of use runs from the south (Vanderbijlpark) to Pretoria in the north.

The result also indicates a steep decline between the Johannesburg central area to the south west in the direction of Westonaria. This steep decline is not evident from the other map.

The circles indicate the position of the sampled police stations and the weighted percentage of arrestees who indicated lifetime alcohol use. These figures correspond very well with the spatial interpolation result.

STATISTICAL COMPARISON

In order to determine which of the methodologies deliver the best result, a correlation was done between the resulting values. This means the spatially interpolated data was assigned to the underlying police stations and this value for lifetime alcohol use was correlated with the neural network value for the same variable. The sampled police stations were excluded from this analysis.

The result of the scatterplot is shown in Figure 5 below. The analyses indicated no correlation between the two methods. It can therefore not be established which method yields the best results.

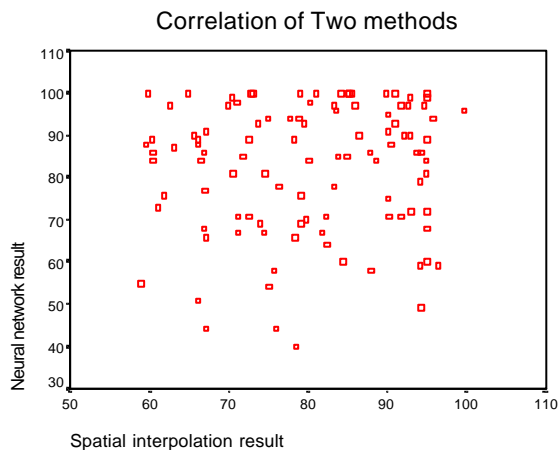


Figure 5

The only way these two can be compared is by going back to the field and to sample a number of police stations that were excluded from the original sample. The extrapolated values of these newly sampled stations should then be compared with the surveyed value for lifetime alcohol use to determine the success rate of the two methods utilised.

WAY FORWARD

It is suggested that a new sample of police stations should be drawn to verify the results of the

extrapolation methods used. Statistical correlation did not yield meaningful results and to implement the spatial interpolation methodology successfully it is important to know the statistical impact.

In terms of spatial comparison the methods yielded similar results for the sampled police stations. It seems therefore as if spatial interpolation can be used successfully.

Two fundamental differences underlie these methodologies however. Neural network back propagation considers multiple factors, like socio-economics together with sampled data, to estimate values for non-sampled areas. Technically speaking it should therefore be able to provide a more holistic picture than spatial interpolation.

The latter is based on the spatial distribution of one variable. This might be a simplification of reality, but poses a solution to surmount problems of complex neural network technology. This simplification should however be taken into consideration in analysis.

The imputed neural network value is basically an extrapolation of observed incidence rates from the underlying socio-economic profile of the unit of spatial analysis. If it is accepted that alcohol usage follows a regional or spatial patterns the contouring of these use patterns can be used in conjunction with the imputed neural network value to reveal discrepancies and anomalies. Large anomalies (differences between expected and imputed rates) indicate regions which bear greater scrutiny – if not in terms of the habits of residents then in terms of the quality of the fieldwork.

REFERENCES

- AGI (Association for Geographic Information). 1999. GIS Dictionary. Internet: <http://www.geo.ed.ac.uk>.
- MacEachren, A.M & Davidson, J.V. 1987. Sampling and Isometric Mapping of Continuous Geographic Surfaces. In *The American Cartographer*, 14, 4: 299-320.
- Singh, R. & Treleaven, P. 1998. Intelligent Systems for GIS. Internet: <http://www.cs.ucl.ac.uk>
- University of British Columbia. 1997. Unit 40 - Spatial interpolation. Internet: <http://www.geog.ubc.ca>.