Geographical General Regression Neural Network (GGRNN) Tool For Geographically Weighted Regression Analysis

Muhammad Irfan, Aleksandra Koj, Hywel R. Thomas, Majid Sedighi
Geoenvironmental Research Centre,
School of Engineering, Cardiff University
Cardiff, UK
emails: {MuhammadI2, KojA, ThomasHR}@cf.ac.uk, majid.sedighi@manchester.ac.uk

***Abstract*—This paper presents a new geographically weighted regression analysis tool, based upon a modified version of a General Regression Neural Network (GRNN). The new Geographic General Regression Neural Network (GGRNN) tool allows for local variations in the regression analysis. The algorithm of the GRNN has been extended to allow for both globally independent variables and local variables, restricted to a given spatial kernel. This mimics the results of Geographically Weighted Regression (GWR) analysis in a given geographical space. The GGRNN tool allows the user to load geographic data from the Shapefile into the underlying neural networks data structure. The spatial kernel can be either a fixed radius or adaptive, by using a given number of neighboring regions. The Holdout Method has been used to compare the fitness of a given model. An application of the tool has been presented using the benchmark working-age deaths in the Tokyo metropolitan area, Japan. Standardized residual maps produced by the GGRNN tool have been compared with those produced by the GWR4 tool for validation. The tool has been developed in the .Net C# programming language using the DotSpatial open source library. The tool is valuable because it allows the user to investigate the influence of spatially non-stationary processes in the regression analysis. The tool can also be used for prediction or interpolation purposes for a range of environmental, socioeconomic and public health applications.**

Keywords- GGRNN; GRNN; GWR; ANN; Spatial Kernel.

I. INTRODUCTION

The GGRNN tool is part of a Spatial Decision Support System (SEREN-SDSS) developed by the Geoenvironmental Research Centre of Cardiff University. SEREN-SDSS has been designed and developed for geoenergy and geoenvironmental applications. It facilitates the decision making process by combining several Multicriteria Decision Analysis (MCDA) and Artificial Neural Network (ANN) techniques [1]. The GGRNN tool utilises and extends the capabilities of GRNN in order to facilitate local spatial variations in regression analyses.

GRNN was first presented by Spetch [2]. GRNN are powerful function approximations, capable of modelling linear and non-linear relationships in data despite being very simple in their structure and operation [3].

GRNNs have been considered in this research because unlike some of the other type of ANNs, GRNNs do not operate as a "black box". Rather, they predict the values at an unknown location on the basis of its proximity to known location in terms of the selected independent variables. Additionally, because of its structure, it is easier to incorporate spatial parameters as one of the independent variables to support local variation in the regression analysis.

The paper is organised as follows. Section II covers the structure, empirical formulation and algorithmic details of the training of the GRNN. Section III describes the nature, operation and different variations of GWR analysis. Section IV presents the GGRNN introduced in this research. Section V highlights the development and operation of the GGRNN tool used here to carry out the GGRNN analysis. Section VI covers the validation of the proposed GGRNN tool. Results obtained using the proposed GGRNN tool, are provided in Section VII together with a comparison of its results against the GWR4 tool. Section VIII summarizes conclusions and future work.

II. GENERAL REGRESSION NEURAL NETWORKS

GRNNs have the capability to predict, interpolate and undertake regression analysis. It is a useful tool when the relationship between dependant and independent variables is unknown and complex. It supports both linear and non-linear relationships.

GRNNs have been used in a number of applications. For example, a GRNN has been used to predict rainwater runoff in two small sub-catchments of Tiber River Basin in Italy using rainfall and soil moisture information at different soil depths [4]. The GRNN prediction was found to be satisfactory in relation to the actual runoff, with coefficient of determination, R_2, equal to 0.87 [4].

Similarly, three different types of neural networks have been used to predict and classify the per-capita Ecological Footprint (EF) of 140 nations [5]. These neural networks are Multi-layer Perceptron Neural Networks (MLPs), Probabilistic Neural Networks (PNNs) and GRNNs. The results reveal that neural networks outperform traditional statistical methods used for this application [5].

GRNNs can also be utilised in finding the most useful set of variables that can be used in an analysis. For example, GRNNs have been used in [6] for the determination of the most appropriate variables to forecast chlorine in preventing the spread of waterborne diseases.

A. Structure of GRNN

GRNNs are very simple in their structure and have the following four layers of neurons: a) Input Layer, b) Pattern Layer, c) Summation Layer, d) Output Layer.

 ISBN: 978-1-61208-469-5

Figure 1 shows the general structure of a GRNN with these four layers, as originally suggested by [2]. A GRNN can approximate a function and estimate the value of a dependent variable from a set of independent variables.

The Input Layer contains as many neurons as there are variables in the input dataset. The input data points are presented to the Input Layer which simply feeds them into the Pattern Layer. Each input data point is then stored in the Pattern Layer. The number of neurons in the Pattern Layer is equal to the total number of data points. The value of the dependent variable (Y) at the prediction point is calculated based on the difference between the values of independent variables at the prediction point and their respective values at other points at which the independent variables are known. The Summation Layer computes the numerator and denominator terms for Equation 1, by using the difference factor of the independent variables (at known and unknown location) and the dependent variable (at known location). The last layer is called the Output Layer where the value of function $\acute{Y} = f(x)$ is computed using (a).

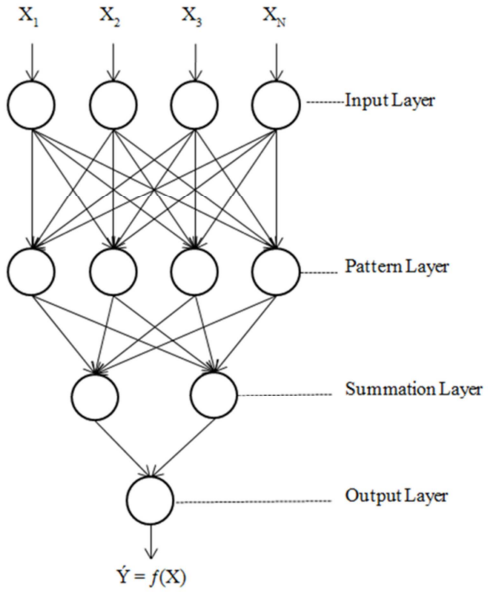

Figure 1. Structure of General Regression Neural Network [2]

The mathematical formulation to implement GRNN is straightforward and similar to probability distribution function. The output function of the GRNN is given as [2]:

$$\acute{Y} = f(x) = \frac{\sum_{i=1}^{n} Y^i \ exp^{(-D_i{}^2/2\sigma^2)}}{\sum_{i=1}^{n} \ exp^{(-D_i{}^2/2\sigma^2)}} \quad (1)$$

where Ý is the estimated value of the dependent variable at the unknown location, Y^i is the value of dependent variable at known locations and D_i is a scalar term that accounts for the differences between the prediction point and the training sample for all independent variables (dimensions) and is calculated as[2]:

$$D_i{}^2 = (X - X^i)^T (X - X^i) \quad (2)$$

The distance between the prediction point and a training sample defines the influence of that training sample in the calculation of $f(x)$ (the dependent variable Ý). If this distance is small, the term $exp^{(-D_i{}^2/2\sigma^2)}$ increases and is exactly one for a difference of zero. A larger value of this term means the known value of dependent variable at this training sample will have more influence in the calculation of the dependant variable at the prediction point. If the distance is large, the value of the term $exp^{(-D_i{}^2/2\sigma^2)}$ decreases, tending to zero for very large distances. Such sample points will provide no contribution to the estimation of dependent variable at the prediction location. The predicted output is bounded between the maximum and minimum known values of the dependent variable [2].

B. Smoothing parameter sigma (σ)

The σ parameter can have single or multiple values for different variables (dimensions) in an input dataset. If a single value is used, it is very important to standardise the independent variables so that they have a mean of zero and a standard deviation of one. Without standardisation of the independent variables, a single σ value will cover different distances in each dimension and the value of $D_i{}^2$ will not represent the actual difference between the training sample and the prediction point [2]. A smaller σ value will result in a localised regression analysis, i.e., only the sample points that are very close to the prediction point in terms of their distances on different axis (domains) will contribute to the calculation of dependent variable. A larger σ value results in a more globalised regression where almost the entire set of data samples contributes to the calculation of the dependent variable. In this latter case, results are very close to the mean value of the dependent variable for the entire set of sample points.

C. Holdout Method for training of GRNN

GRNNs require supervised training and the selection of the most suitable value for the smoothing parameter, σ, is very important to obtain reliable results [2]. The Holdout Method is a useful and common method for the selection of σ [2]. Figure 2 explains the Holdout method algorithm in a flow chart.

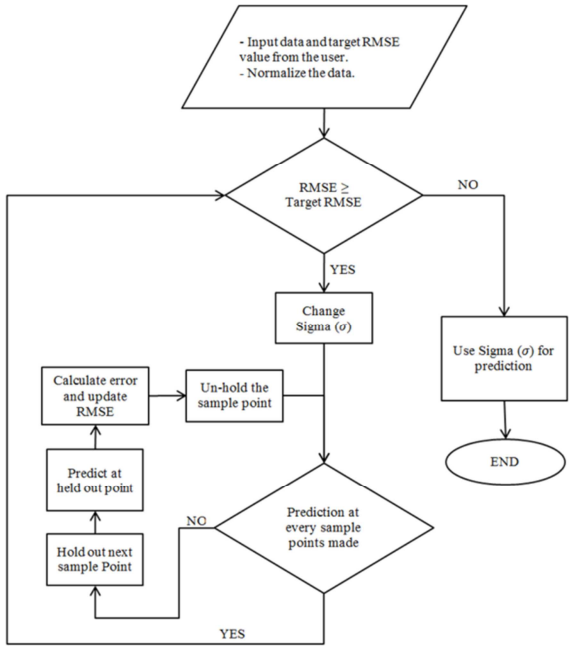

Figure 2. Flow chart of the Holdout Method for the selection of sigma

In the Holdout Method, only one training sample is selected from the training set at a time and the value of Ý is predicted at this sample point using the rest of the samples [2]. The predicted value is compared with the actual value and the difference is used in the calculation of mean squared error [2].

III. GEOGRAPHICALLY WEIGHTED REGRESSION (GWR)

Geographically weighted regression (GWR) models can be used to understand and analyse spatially varying relationships between dependent and independent variables [7]. A conventional GWR regression model is represented by the following equation [7]:

$$Y_i = \sum_k \beta_k(u_i, v_i) x_{k,i} + \varepsilon_i \quad (3)$$

where Y_i, $X_{k,i}$ and ε_i are the dependent variable, *kth* independent variable, and the error term at location *i(u, v)* respectively. β_k is locally varying coefficient at the *i*th location. Another variation of GWR model is where some of the independent are treated as global while others are restricted to vary locally. In such models a user given spatial kernel defines the area in which the local variables are analysed. Such models are called semi-parametric or mixed GWR are normally represented by [7]:

$$Y_i = \sum_k \beta_k (u_i, v_i) x_{k,i} + \varepsilon_i + \sum_l \gamma_l z_{l,i} + \varepsilon_i \quad (4)$$

where $z_{l,i}$ is the l^{th} independent variable that is treated globally and has a fixed coefficient γ_l.

GWR or mixed GWR functions can be applied using Gaussian, Poisson, and logistic regression models. The models give better regression results and enhanced understanding of the relationship between different parameters, whether global or local [7].

IV. GEOGRAPHICAL GENERAL REGRESSION NEURAL NETWORK (GGRNN)

The GGRNN presented in this study extends the basic GRNN model described earlier in Section 2. This extension of the original GRNN algorithm allows for local variation in the relationship between different parameters. The influence of local and global variables are computed separately and then summed together. The difference is in the calculation of the term D (Distance) if spatial distance is used as independent variable as explained earlier. Also for the locally independent variables, the influence is calculated only within the given neighbourhood in contrast to the global variables for which the locations are involved.

In order to define the neighbourhood for local variations, two different techniques are used:

A. *Fixed spatial kernel*

In this technique, a user defined spatial kernel, e.g., 15km, is used to select the neighbouring geographical regions (features). These features are used for the computation of the influence of the local dependent variables only. The influence of global variables is calculated in the normal manner from the entire study area.

B. *Spatially adaptive kernel*

If the spatially adaptive kernel technique is used, the user selects the number of neighbouring areas to define the kernel, within which the influence of the local parameters is calculated. Since the geometries of the administrative boundaries (e.g. districts) are asymmetric, a fixed number of neighbouring areas will result in a varying spatial kernel, hence the naming of this technique.

C. *Spatial distance as independent variable*

The use of an appropriate neighbourhood size is important for the model to fit the data properly. Different iterations and comparison of the standardised error can help in the identification of the appropriate neighbourhood size. However, if it is not clear what type and size of kernel is to be used, the GGRNN tool also provides a mechanism to use spatial distance between different areas as one of the independent variables for the prediction of the dependent variable. As discussed earlier in Section 2, the neighbouring areas of the prediction location will have a greater influence in the calculation of the dependent variable. The distance between two geographical features (areas) is calculated using (5) based on the centroids of either feature [7]:

$$D_{spatial} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2} \quad (5)$$

V. GGRNN TOOL

The GGRNN tool has been developed in the .Net C# programming language using the DotSpatial open source library. Figure 3 shows the user interface of the GRNN based prediction tool.

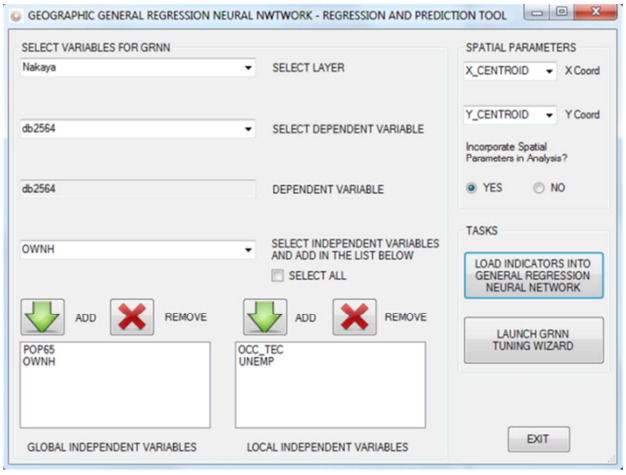

Figure 3. GUI of GGRNN based prediction and regression analysis tool

The user first selects the GIS layer (Shapefile) containing the indicators. The user identifies the dependent, global and local independent variables, and loads the data into the GRNN tool. The user can select whether or not to use spatial distance in the analysis.

 ISBN: 978-1-61208-469-5

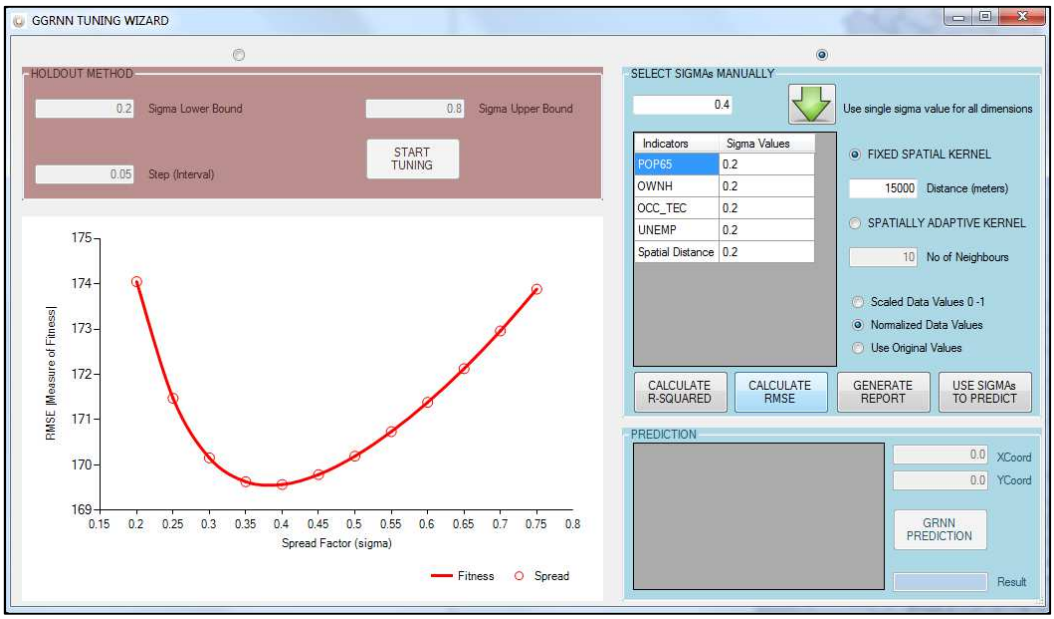

Figure 4. GRNN sigma tuning and prediction tool

Once the data is incorporated in the neural network, a tuning wizard is launched helping the user to select best sigma (σ) parameters for the analysis. The tuning wizard utilises the Holdout Method for the calculation of the Root Mean Square Error (RMSE).

The user can give upper and lower bounds for the sigma parameters and a step (interval) to calculate the RMSE using the Holdout method. The system plots the RMSE values against the corresponding sigma spread factors as shown in Figure 4.

Either the actual, scaled or normalised data values can be used for the calculation of RMSE for a given set of σ values. The user can assign the same σ parameters for all the independent variables if the data is normalised or scaled. However, if the original data values of the independent variables are used for the estimation of the dependent variable, then it is important to assign the sigma values with care. This is important as some of the variables may have a different spread and range of data values as compare to the others and using a similar sigma value can adversely affect the results.

Adopting spatial parameters in the regression analysis in GRNN is similar to the Geographically Weighted Regression (GWR) suggested by [7]. If spatial parameters are included in the analysis then the tool provides two different methods to identify a specific number of neighbouring geographical features to be used for the prediction analysis. These two methods are a) Fixed Spatial Kernel and b) Adaptive Spatial Kernel.

If an Adaptive Spatial Kernel is selected, only a given number (N) of neighbouring geographical features are selected for the analysis. The system first calculates the distance of each geographical feature from the prediction point. Then only N closest neighbours are selected and used in the process. However, if a fixed spatial kernel is used then all neighbouring geographical features found within the spatial kernel are selected.

In either case the smoothing parameter, sigma (σ), used for each independent variable computes the influence of each neighbouring area on the calculation of the independent variable at the prediction point. If a sigma parameter is assigned to the spatial dimension, then features closer to the prediction point will have a greater influence on this calculation. Large values of sigma parameters cause the prediction to tend to the mean value of the dependent variable in the entire study area of the given neighbourhood.

Once a set of sigma parameters has been selected with an acceptable RMSE value, the user can select to use them for the actual prediction at an unknown location. If spatial parameters were not used in the analysis, only the independent variables need to be provided by the user at the unknown location, where prediction is to be made for the dependent variable. If however, spatial parameters were used, then the user must also provide the X and Y coordinated of the centroid of the geographical feature, for which the dependent variable is to be predicted.

VI. VALIDATION

An application of the GGRNN tool is presented to compare its results with those produced by the GWR tool. A semi-parametric GWR model application has been presented to analyse the relationships between the working-age mortality and socio-economic conditions in Tokyo metropolitan area, Japan [8]. The same dataset is used in this research for two reasons:

a. The dataset is known to have local spatial variations found in parts of the study area, as explained in [8].
b. The standardised error resulting in the application of GWR and the GGRNN tool can be mapped, analysed and compared for benchmarking purpose.

The Tokyo mortality data covers the 262 municipality zones of the Tokyo Metropolitan area, Japan. The older age population and rate of house-ownership are used by

 ISBN: 978-1-61208-469-5

[8] as the global independent variables, whereas the other two variables are controlled locally in the regression analysis. The description of dependent and independent variables are given in Table 1 below.

TABLE 1. TOKYO MORTALITY DATASET

Variable	Description	Relationship
Working age mortality rate	Standard mortality rates for the 25–64 age group	Dependent Variable
Older population	Proportion of elderly people (aged over 64) within each zone	Independent (Global)
Own houses	Rate of house-ownership in each zone	Independent (Global)
Professional and technical workers	Proportion of professional and technical workers in each zone.	Independent (Local)
Unemployment	Rate of unemployment in each zone	Independent (Local)

VII. COMPARISON OF RESULTS

GWR version 4 has been used to analyse the Geographically Weighted Regression of working age mortality rates with socio economic conditions. A Gaussian Model has been used for the kernel analysis in both the GWR and GGRNN tools. The introduction of an offset and a local intercept variable in the GWR analysis is recommended [8]. Therefore, the two variables have been included in the GWR tool; however, the GGRNN tool doesn't have a provision for this because of the structure of its underlying neural network. In both cases the independent variables are standardised. Both fixed and adaptive kernels have been used to run the model in GWR. The recommended fixed kernel for this dataset is 15km and, for an adaptive kernel type, 50 neighbours are recommended [8]. In order to compare the results with those produced by the GGRNN tool, the most suitable sigma parameter is identified using the Holdout Method and RMSE. A sigma value of 0.4 was obtained for both adaptive and fixed spatial kernel techniques. Standardised residual maps are produced in ArcMap; the resultant maps obtained using the GWR tool, are shown in Figure 5 below.

Figure 5 shows the standardized residual maps produced by the GGRNN tool and GWR4 tool by using an adaptive kernel. The results show that the GGRNN tool has produced very similar results to the GWR4 tool using the adaptive kernel. A slight difference can be observed between the two results in the south-eastern part of the region which needs to be further investigated. A possible reason is the difference between the locally varying coefficient used in the GWR tool and the sigma parameter used in the GGRNN tool.

In the second process, both the tools have been set to use a fixed spatial kernel of 15 km. The Holdout Method used in the GGRNN tool suggests that a network model with sigma parameter of 0.4 exhibits the best fit to the dataset. The results are shown in Figure 6. It can be seen that the two tools have again produced very similar results in the case of fixed spatial kernel.

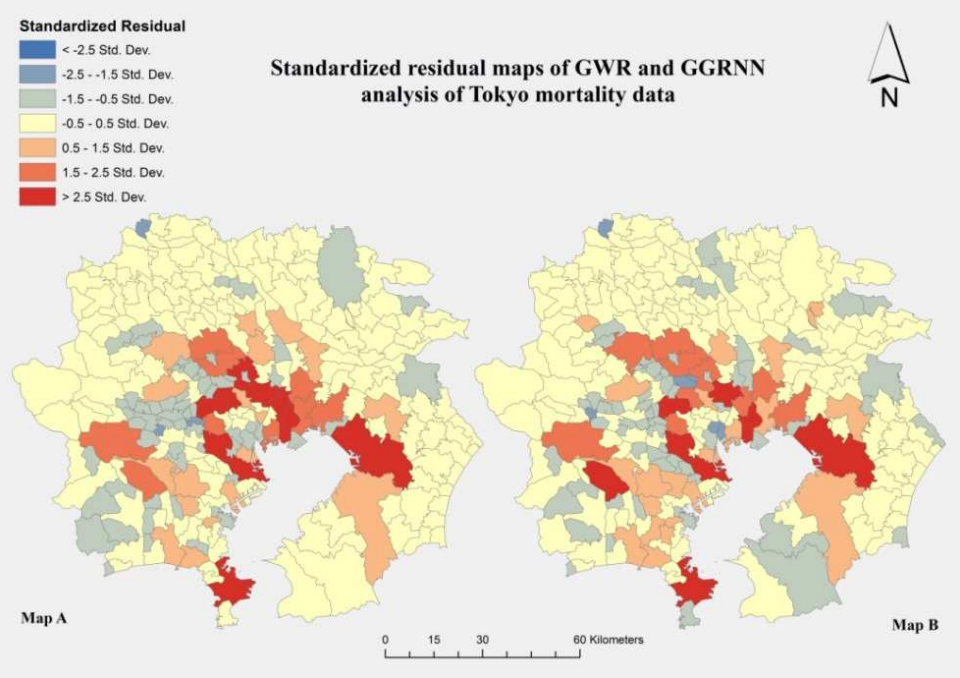

Figure 5. Standardised residual maps using adaptive Gaussian with 50 neighbours. Map A: GGRNN tool. Sigma parameter: 0.4 (for all independent variables). Map B: GWR4 tool

 ISBN: 978-1-61208-469-5

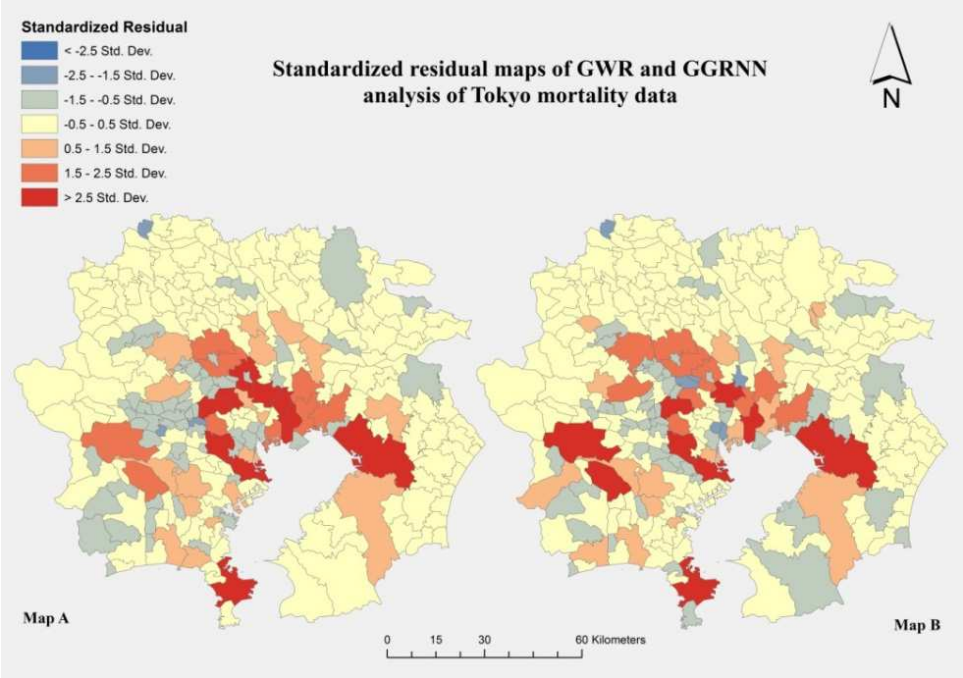

Figure 6. Standardised residual maps using fixed kernel of 15kms. Map A: GGRNN tool. Sigma parameter: 0.4 (for all independent variables). Map B: GWR4 tool

VIII. CONCLUSIONS AND FUTURE WORK

This paper presents a new regression analysis tool, based upon a modified version of the General Regression Neural Network (GRNN). The Geographical General Regression Neural Network (GGRNN) tool can be used to perform to Geographically Weighted Regression (GWR) analysis. It can be useful in understanding the underlying spatially varying relationships between dependent and independent variables and for prediction analysis. The GGRNN tool can be used in a number of environmental, socio-economic and public health applications.

The tool provides options to select the independent variables as globally fixed or locally varying. The spatial kernel can either be assigned as a fixed radius or adaptive, i.e., by assigning a given number of neighbouring regions.

The Holdout Method has been used to compare the fitness of a given model. The GGRNN tool allows the user to compare the fitness of different models by using the Holdout Method. The Holdout Method helps in selecting the most appropriate network parameters, essential for the working of a neural network. A validation of the tool has been carried out using the benchmark Tokyo mortality dataset and using the GWR4 tool. The validation results demonstrate that the GGRNN tool can be used with confidence to carry out geographically weighted regression analysis.

In future work, the performance of the tool will be tested against the GWR tool. Also, it will be tested to assess its prediction of dependent variable at unknown locations for impact assessment.

ACKNOWLEDGMENT

The work described in this paper has been carried out as part of the GRC's Seren project (GRC SEREN Project 2015) [9], which is part funded by the Welsh European Funding Office (WEFO).

REFERENCES

[1] M. Irfan, "An Integrated, Multicriteria, Spatial Decision Support System, Incorporating Environmental, Social and Public Health Perspectives, for Use in Geoenergy and Geoenvironmental Applications", Ph.D. Thesis, Cardiff University, The Wales, UK (2014).

[2] D. F. Specht, "A General Regression Neural Network", Neural Networks, IEEE Transactions on, vol. 2, 1991, pp. 568-76.

[3] N. Currit, "Inductive Regression: Overcoming OLS Limitations with the General Regression Neural Network", Computers, Environment and Urban Systems, vol. 26 2002, pp. 335-53.

[4] G. Tayfur, G. Zucco, L. Brocca, and T. Moramarco, "Coupling Soil Moisture and Precipitation Observations for Predicting Hourly Runoff at Small Catchment Scale", Journal of Hydrology, vol. 510, 2014, pp. 363-71.

[5] M. M. Mostafa, and R. Nataraajan, "A Neuro-Computational Intelligence Analysis of the Ecological Footprint of Nations", Computational Statistics & Data Analysis, vol. 53, 2009, pp. 3516-31.

[6] G. J. Bowden, J. B. Nixon, G. C. Dandy, H. R. Maier, and M. Holmes, "Forecasting Chlorine Residuals in a Water Distribution System Using a General Regression Neural Network", Mathematical and Computer Modelling, vol. 44, 2006, pp. 469-84.

[7] S. Fotheringham, C. Brunsdon, and M. Charlton, Geographically Weighted Regression: The Analysis of Spatially Varying Relationships (Wiley, 2003).

[8] T. Nakaya, A. S. Fotheringham, C. Brunsdon, and M. Charlton, "Geographically Weighted Poisson Regression for Disease Association Mapping", Statistics in Medicine, vol. 24, 2005, pp. 2695-717.

[9] GRC SEREN Project. [online] Available from: http://grc.engineering.cf.ac.uk/research/seren/ 2016.04.14

 ISBN: 978-1-61208-469-5