Math Geosci (2010) 42: 657–680
DOI 10.1007/s11004-010-9284-7

The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets

P. Harris · A.S. Fotheringham · R. Crespo · M. Charlton

Received: 7 January 2010 / Accepted: 25 April 2010 / Published online: 11 June 2010
© International Association for Mathematical Geosciences 2010

Abstract Increasingly, the geographically weighted regression (GWR) model is being used for spatial prediction rather than for inference. Our study compares GWR as a predictor to (a) its global counterpart of multiple linear regression (MLR); (b) traditional geostatistical models such as ordinary kriging (OK) and universal kriging (UK), with MLR as a mean component; and (c) hybrids, where kriging models are specified with GWR as a mean component. For this purpose, we test the performance of each model on data simulated with differing levels of spatial heterogeneity (with respect to data relationships in the mean process) and spatial autocorrelation (in the residual process). Our results demonstrate that kriging (in a UK form) should be the preferred predictor, reflecting its optimal statistical properties. However the GWR-kriging hybrids perform with merit and, as such, a predictor of this form may provide a worthy alternative to UK for particular (non-stationary relationship) situations when UK models cannot be reliably calibrated. GWR predictors tend to perform more poorly than their more complex GWR-kriging counterparts, but both GWR-based models are useful in that they provide extra information on the spatial processes generating the data that are being predicted.

Keywords Relationship nonstationarity · Relationship heterogeneity · GWR · Kriging · Spatial interpolation

P. Harris (✉) · A.S. Fotheringham · M. Charlton
National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland
e-mail: Paul.Harris@nuim.ie

R. Crespo
Institute for Spatial and Landscape Planning, Swiss Federal Institute of Technology, Zurich, Switzerland

1 Introduction

Spatial prediction is of great interest in many areas of applied science, and has an extensive history of development. The ability to predict values at unknown locations is useful not only for scientists who attempt to model spatial processes, but also for policy makers who need to plan and manage the outcomes of spatial processes at regional or local levels. Despite the longevity of the subject matter, there is always an element of uncertainty concerning what type of prediction method is most appropriate in different situations. Much depends on the nature of the sample data and the decisions made by the analyst when parameterising the predictor. Often the concept of external objectivity (Matheron 1989, p. 38) is invoked, for which the worth of a given predictor can be gauged by its performance in the long run through an increasing number of applications.

Kriging (Matheron 1963, 1969) is a best linear unbiased predictor, and this statistical property ensures that it is frequently preferred. To allow for different ways of modelling the mean (or trend) in kriging, it is common to decompose the spatial process into a mean and residual component, where all the randomness of the data is associated with the residual component. For ordinary kriging (OK), the mean component is taken as some constant, whereas for universal kriging (UK), it is taken as some non-constant, such as that found with a multiple linear regression (MLR) fit. The MLR fit can be a function of the coordinates in the univariate case, or a function of external covariates in the multivariate case. In the latter instance, UK is often re-named kriging with external drift (KED) or regression kriging (RK) depending on whether an implicit (mean and residual components found simultaneously in a single-stage procedure) or explicit (mean and residual components found separately in a two-stage procedure) solution to the UK model is adopted, respectively (Hengl et al. 2007). Often the mean component of UK is just as accurate as UK itself, reflecting a residual process that exhibits little or no spatial autocorrelation. Here, the simpler MLR component fit can be preferred. Thus in our study, standard predictors such as MLR, OK and UK are calibrated and compared to the geographically weighted regression (GWR) model (Fotheringham et al. 2002), in which the influence of explanatory data is modelled locally via many localised MLR fits. As with our MLR and UK models, our GWR models are only calibrated using external covariate data.

Furthermore, kriging can be approximated with the use of local neighbourhoods (i.e. only using data that are close to target locations), which in the implicit form of UK allows the MLR component fit to be calibrated locally. In this respect, we calibrate this particular UK model to act as a standard geostatistical alternative to GWR, as it can similarly account for relationship nonstationarity between a dependent variable and its covariates. This UK specification can be viewed as an implicit GWR-kriging hybrid, since its mean component is GWR specified with a box-car kernel. Similarly, it is also possible to specify an explicit GWR-kriging hybrid, where the GWR mean component can be more generally defined with any kernel function (e.g. a distance-decay, exponential kernel, etc.). In this kriging construction, the mean component is globally nonparametric and globally non-linear. Kriging studies have experimented with nonparametric (e.g. Cressie 1986; Genton and Furrer 1998; Kanevski and Maignan 2004) and non-linear (e.g. Gambolati and Volpi 1979;

Neuman and Jacobson 1984; Haas 1996) mean components before, and in this respect, we also aim to investigate the value of this particular GWR-kriging hybrid (named GWRK, which acts as our third non-stationary relationship model). Again, if a GWR mean component is just as accurate as GWRK itself, the simpler GWR fit should be preferred.

The use of GWR as a predictor has only recently attracted attention (e.g. the empirical evaluations of Zhang et al. 2005; Gao et al. 2006; Bitter et al. 2007). However, only Páez et al. (2008) and Lloyd (2010) are known to specifically compare GWR to kriging in this respect. In the former study, GWR performed favourably against UK (in its full, global neighbourhood form) and against a UK-based moving window kriging (MWK) model (local neighbourhoods and local spatial autocorrelation, as in Haas 1990, 1996), when predicting house price data. In the latter study, a more extensive collection of GWR and kriging models were investigated, including GWR-kriging hybrids, some of which are similar in construction to that used in this study. Here a UK-based MWK model performed the best of all, when predicting monthly precipitation across the United Kingdom. Our study now adds to such literature where we compare the prediction performance of five core techniques (MLR, GWR, OK, UK, and GWRK), highlighting the utility (if any) of those techniques based on GWR via simulated data comparisons. Here we not only assess prediction accuracy, but, unlike previous studies, we also assess estimates of prediction uncertainty accuracy.

2 Prediction Techniques

All models can be defined using $Z(\mathbf{x}) = m(\mathbf{x}) + R(\mathbf{x})$, where the random function $Z(\mathbf{x})$ is decomposed into a mean $m(\mathbf{x})$ and residual $R(\mathbf{x})$ component. Here $R(\mathbf{x})$ describes fluctuations about the mean (i.e. second-order variation); $\mathbf{x}$ is any spatial location (observed or unobserved); and $z(\mathbf{x}_i)$ is the data with $i = 1, \ldots, n$. MLR and GWR model $m(\mathbf{x})$ assuming that $R(\mathbf{x})$ is a stationary random function with $\mathrm{E}\{R(\mathbf{x})\} = 0$ and $\mathrm{VAR}\{R(\mathbf{x})\} = \boldsymbol{\Sigma}$, where the elements of the diagonal $(n \times n)$ matrix $\boldsymbol{\Sigma}$ reflect a pure nugget covariogram (i.e. no spatial autocorrelation, with $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$). OK, UK, and GWRK each model $m(\mathbf{x})$ as some constant, some MLR fit or some GWR fit, respectively. However, for kriging, the elements of $\boldsymbol{\Sigma}$ reflect a structured covariogram $C(\mathbf{h})$, where $\mathbf{h}$ is the separation distance vector $\mathbf{h} = \mathbf{x}_i - \mathbf{x}_j$ (i.e. spatial autocorrelation exists). As is standard practice, the elements of $\boldsymbol{\Sigma}$ are found from the variogram $\gamma(\mathbf{h})$ and using the relationship $C(\mathbf{h}) = \sigma^2 - \gamma(\mathbf{h})$. Accordingly, $\boldsymbol{\Sigma}$ is a function of variogram parameters and can be denoted by $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, where for this study, $\boldsymbol{\theta}$ is a variogram parameter vector consisting of a (small-scale) nugget variance c_0; a (large-scale) structural variance c_1 (where $c_0 + c_1 = \sigma^2$); and a correlation range a.

2.1 MLR and GWR

For the case where there are several independent covariates $y_1, y_2, \ldots, y_k$, the MLR model can be written as $\mathbf{Z} = \mathbf{Y}\boldsymbol{\beta} + \mathbf{R}$, where $\mathbf{Z}$ is the $(n \times 1)$ sample (dependent) data vector, $\mathbf{Y}$ is the $(n \times (k+1))$ covariate matrix, $\boldsymbol{\beta}$ is a $((k+1) \times 1)$ vector of unknown parameters, and $\mathbf{R}$ is a $(n \times 1)$ residual vector. Here the ordinary least-squares (OLS)

parameter estimates $\hat{\boldsymbol{\beta}}$ are found from $\hat{\boldsymbol{\beta}} = (\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{Z}$, and the MLR prediction at $\mathbf{x}$ (where $\mathbf{y}(\mathbf{x})$ is a $((k+1)\times 1)$ vector of covariates at $\mathbf{x}$) is

$$\hat{z}_{\mathrm{MLR}}(\mathbf{x}) = \mathbf{y}(\mathbf{x})^{\mathrm{T}}\hat{\boldsymbol{\beta}}. \tag{1}$$

The corresponding GWR model results in the (location-specific) parameter estimates $\hat{\boldsymbol{\beta}}(\mathbf{x})$ being found from $\hat{\boldsymbol{\beta}}(\mathbf{x}) = (\mathbf{Y}^{\mathrm{T}}\mathbf{W}(\mathbf{x})\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{W}(\mathbf{x})\mathbf{Z}$ where $\mathbf{W}(\mathbf{x})$ is a $(n\times n)$ diagonal matrix of spatial weights (i.e. parameters are estimated using weighted least squares (WLS) with weights changing according to location $\mathbf{x}$). The GWR prediction at $\mathbf{x}$ is

$$\hat{z}_{\mathrm{GWR}}(\mathbf{x}) = \mathbf{y}(\mathbf{x})^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\mathbf{x}). \tag{2}$$

MLR and GWR prediction variances at $\mathbf{x}$, $\sigma^2_{\mathrm{MLR}}(\mathbf{x})$ and $\sigma^2_{\mathrm{GWR}}(\mathbf{x})$, are estimated using

$$\mathrm{VAR}\left[\hat{z}(\mathbf{x}) - z(\mathbf{x})\right] = \hat{\sigma}^2\left[1 + \mathrm{S}(\mathbf{x})\right]. \tag{3}$$

Here, an unbiased estimate of the residual variance is taken as $\hat{\sigma}^2 = \mathrm{RSS}/(n - \mathrm{ENP})$, where RSS is the residual sum of squares and ENP is the effective number of parameters of the MLR or GWR model. Further, for MLR

$$\mathrm{S}(\mathbf{x}) = \mathbf{y}(\mathbf{x})^{\mathrm{T}}\left[\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right]^{-1}\mathbf{y}(\mathbf{x}), \tag{4}$$

and for GWR

$$\mathrm{S}(\mathbf{x}) = \mathbf{y}(\mathbf{x})^{\mathrm{T}}\left[\mathbf{Y}^{\mathrm{T}}\mathbf{W}(\mathbf{x})\mathbf{Y}\right]^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{W}^2(\mathbf{x})\mathbf{Y}\left[\mathbf{Y}^{\mathrm{T}}\mathbf{W}(\mathbf{x})\mathbf{Y}\right]^{-1}\mathbf{y}(\mathbf{x}), \tag{5}$$

respectively. For details, use Leung et al. (2000).

In this study, the weighting matrix in GWR is specified using either a (discontinuous) box-car or (continuous) exponential kernel. The use of a box-car kernel results in the simpler, moving window regression model, where weights at location $\mathbf{x}$ accord to $w(\mathbf{x}) = 1$ if $d_i \le \tau$ and $w(\mathbf{x}) = 0$ if $d_i > \tau$. Here the bandwidth parameter is the distance τ and d_i is the distance between $\mathbf{x}$ and a sample location i. An exponential kernel is defined as $w(\mathbf{x}) = \exp(-d_i/b)$ where the bandwidth parameter is the distance b. For this study, an optimal bandwidth is found in an adaptive form (i.e. the bandwidth reflects a fixed local sample density instead of a fixed distance) using two techniques.

Firstly, it is found using leave-one-out cross-validation, where the root mean squared error (RMSE) is calculated for a range of bandwidths and the bandwidth that gives the minimum RMSE is considered optimal. At each cross-validation point, $\mathrm{RMSE} = \sqrt{(1/(n-1))\sum_{i=1}^{n-1}\{z(\mathbf{x}_i) - \hat{z}(\mathbf{x}_i)\}^2}$, where $\hat{z}(\mathbf{x}_i)$ is the GWR prediction at sample point i when $z(\mathbf{x}_i)$ is omitted from the computation. Secondly, it is found using the (bias corrected) AIC procedure outlined in Fotheringham et al. (2002, p. 61, p. 96), as there is a risk that a GWR model using an RMSE-defined bandwidth may over-fit the calibration data, resulting in poor model information when it comes to prediction at validation sites. Adopting this second technique allows us to gauge whether or not there are instances of over-fitting in our simulation experiment. Here the bandwidth that results in the smallest model AIC is considered optimal. An AIC approach

is ideal to investigate over-fitting issues with GWR, as it penalises model complexity, providing a bandwidth that reducing instances of under-smoothing (i.e. AIC-defined bandwidths will tend to be larger than RMSE-defined ones). Throughout this study, all bandwidths are presented as a percentage, which for a box-car kernel relates to local sample size, whilst for an exponential kernel reflects a local sample size that exerts the greatest influence on each local regression fit.

2.2 OK and UK

Here we briefly outline the standard UK model where $m(\mathbf{x})$ is modelled using MLR. The OK model is not presented, as it is a straightforward simplification of UK where $m(\mathbf{x})$ is modelled as an unknown constant (e.g. Schabenberger and Gotway 2005, p. 241–243). For unbiased variogram estimation in UK, it is necessary to find the variogram of the residual process $\mathbf{Z} - \mathbf{Y}\boldsymbol{\beta}$. However, $\boldsymbol{\beta}$ is unknown and can only be estimated efficiently with generalised least squares (GLS) which itself needs to be calibrated using unbiased variogram information, via $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. This is the well-known analytical impasse of UK, which is commonly addressed via the use of a restricted (or residual) maximum likelihood (REML) approach to first identify relatively unbiased estimates of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ and then in turn, relatively unbiased estimates of $\boldsymbol{\beta}$ (e.g. Schabenberger and Gotway 2005, pp. 259–263). Such an approach is adopted in this study to parameterise both OK and UK, where only exponential variogram models are considered, i.e. $\gamma(h) = c_0 + c_1(1 - \exp(-h/a))$.

Thus, for UK, $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\mathbf{Y}^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\mathbf{Z}$ provides the (best linear unbiased) parameter estimates and the (best linear unbiased) prediction at $\mathbf{x}$ is

$$\hat{z}_{\mathrm{UK}}(\mathbf{x}) = \mathbf{y}(\mathbf{x})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} + \boldsymbol{\sigma}_{\boldsymbol{\theta}}^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}(\mathbf{Z} - \mathbf{Y}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}). \tag{6}$$

Here $\boldsymbol{\sigma}_{\boldsymbol{\theta}}$ is a $(n \times 1)$ vector of spatial covariances between residuals at $\mathbf{x}$ and the sample locations. The UK weights at $\mathbf{x}$ are $\boldsymbol{\lambda}_{\mathrm{UK}}(\mathbf{x}) = [\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\boldsymbol{\sigma}_{\boldsymbol{\theta}}$ and the UK variance at $\mathbf{x}$ is

$$\begin{aligned}\sigma^2_{\mathrm{UK}}(\mathbf{x}) = {} & \left[\hat{\sigma}^2 - \boldsymbol{\sigma}_{\boldsymbol{\theta}}^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\boldsymbol{\sigma}_{\boldsymbol{\theta}}\right] \\ & + \left[\mathbf{y}(\mathbf{x}) - \mathbf{Y}^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\boldsymbol{\sigma}_{\boldsymbol{\theta}}\right]^{\mathrm{T}}\left[\mathbf{Y}^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\mathbf{Y}\right]^{-1} \\ & \times \left[\mathbf{y}(\mathbf{x}) - \mathbf{Y}^{\mathrm{T}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\boldsymbol{\sigma}_{\boldsymbol{\theta}}\right].\end{aligned} \tag{7}$$

The first part of (7) represents the kriging variance of the residuals and the second part is a consequence of estimating the trend with MLR. Observe that $\hat{\sigma}^2$ is the estimate of the residual variogram sill $(c_0 + c_1)$, where c_0 and c_1 are partial sills.

To find $\hat{z}_{\mathrm{UK}}(\mathbf{x})$ and $\sigma^2_{\mathrm{UK}}(\mathbf{x})$ both implicit and explicit solutions are possible, although these are identical if a global neighbourhood is specified (Chilès and Delfiner 1999; Rivoirard 2002; Hengl et al. 2007). In this study, an implicit solution is adopted which allows the MLR component to be calibrated locally when UK is specified with local neighbourhoods. A neighbourhood size is optimally found using the same cross-validation procedure as that used to find the bandwidth in GWR. Technically, local residual variogram parameters should now be estimated that are specific to $\mathbf{x}$

(as in a MWK model), but instead, the parameters of the globally-found residual variogram are retained and this pragmatic modelling decision is usually referred to as a quasi-stationarity (Journel and Huijbregts 1978, pp. 33–34). When kriging is approximated in this way, it is no longer a best linear unbiased predictor (Chilès and Delfiner 1999, p. 201) and poorly chosen neighbourhoods can result in unwanted discontinuities. However, the approximation usually has little effect on overall prediction accuracy and is routinely used to ease computational burden with large data sets. For this study, the approximation is specifically used to model non-stationary relationships via the UK model as it provides an obvious means to improve prediction over a stationary relationship counterpart. It is not so easy to justify the approximation from a prediction uncertainty viewpoint as the kriging variance depends on the variogram, except that the reliable estimation of local variogram parameters is commonly fraught with technical difficulties (e.g. Atkinson 2001; Schabenberger and Gotway 2005, pp. 425–426).

2.3 GWRK

Unlike the implicit solution, the explicit solution to UK deals with the mean and residual processes separately in a distinct two-stage procedure where the mean component is found first and then kriging is performed on the residual data. Only the explicit approach lends itself to nonparametric or non-linear mean components; therefore, it is adopted for our hybrid GWRK model. Optimal parameterisation via REML is not viable for GWRK and as such, the GWR mean component is found as described in Sect. 2.1, and the corresponding residual variogram $\gamma_R(\mathbf{h})$ is estimated using the usual classical estimator (e.g. Schabenberger and Gotway 2005, pp. 153–154). This estimator is then modelled with an exponential variogram using the WLS variogram fitting approach of Zhang et al. (1995).

As with many explicit approaches, GWRK has a focus on an efficient estimation of the trend and accepts that some residual variogram bias is ever-present. In this respect, explicit approaches are not always built on any theoretical basis, but commonly provide satisfactory results. In terms of residual variogram bias, nonparametric/non-linear trend fits that are strongly local can capture most of the variation in the data resulting in a residual variogram that is either pure nugget or if structured, is highly biased (much more so than those found with an equivalent linear trend, see Pardo-Igúzquiza and Dowd 1998). As such, a GWRK prediction at $\mathbf{x}$ is defined as

$$\hat{z}_{\mathrm{GWRK}}(\mathbf{x}) = \hat{z}_{\mathrm{GWR}}(\mathbf{x}) + \hat{r}_{\mathrm{OK}}(\mathbf{x}), \tag{8}$$

if the residual variogram is structured, and $\hat{z}_{\mathrm{GWRK}}(\mathbf{x}) = \hat{z}_{\mathrm{GWR}}(\mathbf{x})$ if the residual variogram is modelled as a pure nugget variance. Observe that we have specified OK with the residual data. If simple kriging (SK) were specified instead, then instances of $\hat{z}_{\mathrm{GWRK}}(\mathbf{x}) = \hat{z}_{\mathrm{GWR}}(\mathbf{x})$ would be directly guaranteed as SK with a pure nugget variogram yields zero predictions.

GWRK variances are not so easily found via a simple addition, since (a) there are complex correlations between the two component processes and (b) the residual variance estimate $\hat{\sigma}^2_{\mathrm{GWR}}$ used in the GWR prediction variance calculation is unlikely to correspond to the residual variogram sill estimate $\hat{\sigma}^2_{\mathrm{SILL}}$ (it is commonly biased

downwards) used in the residual OK variance calculation. If the two residual variance estimates in (b) were equal, then an approximate GWRK variance at $\mathbf{x}$ could be taken as $\sigma^2_{\mathrm{GWRK}}(\mathbf{x}) = \{\hat{\sigma}^2 - \boldsymbol{\sigma}^{\mathrm{T}}_{\boldsymbol{\theta}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\boldsymbol{\sigma}_{\boldsymbol{\theta}}\} + \{\hat{\sigma}^2 \mathrm{S}(\mathbf{x})\}$, where the first part of this expression represents the OK variance of the GWR residuals and the second part is a consequence of estimating the trend with GWR (i.e. analogous to the UK composite variance in (7)). Instead, a pragmatic GWRK variance approximation is calculated, which we define as

$$\sigma^2_{\mathrm{GWRK}}(\mathbf{x}) = \hat{\sigma}^2_{\mathrm{GWR}} + \hat{\sigma}^2_{\mathrm{GWR}}\mathrm{S}(\mathbf{x}) + \hat{\sigma}^2_{\mathrm{SILL}} - \boldsymbol{\sigma}^{\mathrm{T}}_{\boldsymbol{\theta}}[\boldsymbol{\Sigma}_{\boldsymbol{\theta}}]^{-1}\boldsymbol{\sigma}_{\boldsymbol{\theta}} - \hat{\sigma}^2_{\mathrm{SILL}}, \tag{9}$$

where the first part of this expression is the GWR prediction variance $\sigma^2_{\mathrm{GWR}}(\mathbf{x})$ (see (3) and (5)), the second part reflects the residual OK variance (see (7)), and the third part is the residual variogram sill estimate. Here $\hat{\sigma}^2_{\mathrm{SILL}}$ is actually fixed to equal the sample residual variance for the WLS variogram fit, so that consistent GWRK variances are found. If the residual variogram is modelled as a pure nugget variance, then the GWRK variance at $\mathbf{x}$ is taken as $\sigma^2_{\mathrm{GWRK}}(\mathbf{x}) = \sigma^2_{\mathrm{GWR}}(\mathbf{x})$.

2.4 Model Summary

In summary, twelve models are calibrated and assessed. These are OK, MLR, UK-GN (global neighbourhood), UK-LN (local neighbourhood), GWR-BX-CV (box-car kernel and bandwidth found by cross-validation), GWR-EXP-CV (exponential kernel and bandwidth found by cross-validation), GWR-BX-AIC (box-car kernel and bandwidth found by AIC), GWR-EXP-AIC (exponential kernel and bandwidth found by AIC), GWRK-BX-CV (GWR-BX-CV as a mean component), GWRK-EXP-CV (GWR-EXP-CV as a mean component), GWRK-BX-AIC (GWR-BX-AIC as a mean component), and GWRK-EXP-AIC (GWR-EXP-AIC as a mean component). OK is the only univariate model, whereas all other models use covariate data to inform the mean component. OK and OK of the residuals in GWRK are specified with a global neighbourhood. Observe that UK-LN is commonly named a KED model and UK-GN is commonly named an RK model. Observe also that the difference between UK-LN and any GWRK model is a subtle one. Both model constructions use (a) a local trend; (b) a global residual variogram; and (c) some form of approximation. However UK-LN follows an implicit solution, whilst GWRK follows an explicit solution.

3 Simulated Data

Although empirical data comparisons are useful when evaluating predictors, results always depend upon the particular properties of the data set used. Sample size, sample configuration, sample variation, distribution shape, spatial heterogeneity and spatial autocorrelation are just some of the many factors that determine the utility of a given predictor. In this respect, it is useful to simulate data sets with known properties and then assess predictors according to these data. Thus we simulate data sets to solely investigate issues of spatial heterogeneity (with respect to data relationships in the mean process) and spatial autocorrelation (with respect to the residual process). In particular, we follow a hybrid simulation approach based on methods used in Farber

and Páez (2007) and Wang et al. (2008) for investigating different GWR models, and Zimmerman et al. (1999) when comparing inverse distance weighting (IDW) to UK. A key impasse to using simulated data is that the method used is often based on a prediction method under scrutiny and as a consequence, this same method tends to perform unfairly better than others. This drawback has to be borne in mind when interpreting the results.

3.1 The Simulation Algorithm

The spatial layout for the simulation is a square region with side lengths of 12 units. To this region, a coordinate system is built with its origin in the bottom left-hand corner. Here 625 observation points are located on a 25×25 lattice with a distance of 0.5 between any two horizontally or vertically neighbouring points. The spatial coordinates of the locations (u_i, v_i) are calculated using

$$(u_i, v_i) = \left[0.5 \bmod \left[(i-1)/25\right], 0.5 \operatorname{int}\left[(i-1)/25\right]\right] \quad \text{for } i = 1, 2, \ldots, 625, \tag{10}$$

where $\bmod((i-1)/25)$ is the remainder of $(i-1)$ divided by 25 and $\operatorname{int}((i-1)/25)$ is the integer part of the number $(i-1)/25$.

The data generating process is then defined as

$$z_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i) y_i + r_i \quad \text{for } i = 1, 2, \ldots, 625, \tag{11}$$

where $\beta_0(u_i, v_i) + \beta_1(u_i, v_i) y_i$ represents the mean process $m(\mathbf{x})$ and r_i represents the residual process $R(\mathbf{x})$ of the decomposed model, $Z(\mathbf{x}) = m(\mathbf{x}) + R(\mathbf{x})$. The y_i observations of the single independent covariate are randomly drawn from a uniform distribution over the interval $(0, 1)$. The two regression coefficients $\beta_0(u_i, v_i)$ and $\beta_1(u_i, v_i)$ are found according to the following three cases.

- Case 1, zero heterogeneity: $\beta_0(u, v) = 1$, $\beta_1(u_i, v_i) = 6.5$.
- Case 2, low heterogeneity: $\beta_0(u_i, v_i) = 1 + (1/6)(u + v)$, $\beta_1(u_i, v_i) = u/3$.
- Case 3, high heterogeneity: $\beta_0(u_i, v_i) = 1 + 4\sin[(1/12)\pi u]$,

$$\beta_1(u_i, v_i) = 1 + (1/324)\left[36 - (6-u)^2\right]\left[36 - (6-v)^2\right].$$

For cases 2 and 3, the regression coefficient surfaces are depicted in Fig. 1. Case 2 represents a fairly simple, non-stationary relationship between z_i and y_i, whilst case 2 represents a more complex, non-stationary relationship.

The resultant mean processes that are generated do not result in the same levels of variation and, as such, all corresponding residual processes are generated at a fixed proportion to the levels of variation found for each mean process. Since experience suggests that the mean component tends to dominate a spatial process, data generation proceeds such that 66.67% of the variation in the dependent data z_i is explained by the mean component (further investigations could vary this proportion). With this constraint in place, the residual terms r_i are generated for each heterogenic case according to the following three cases.

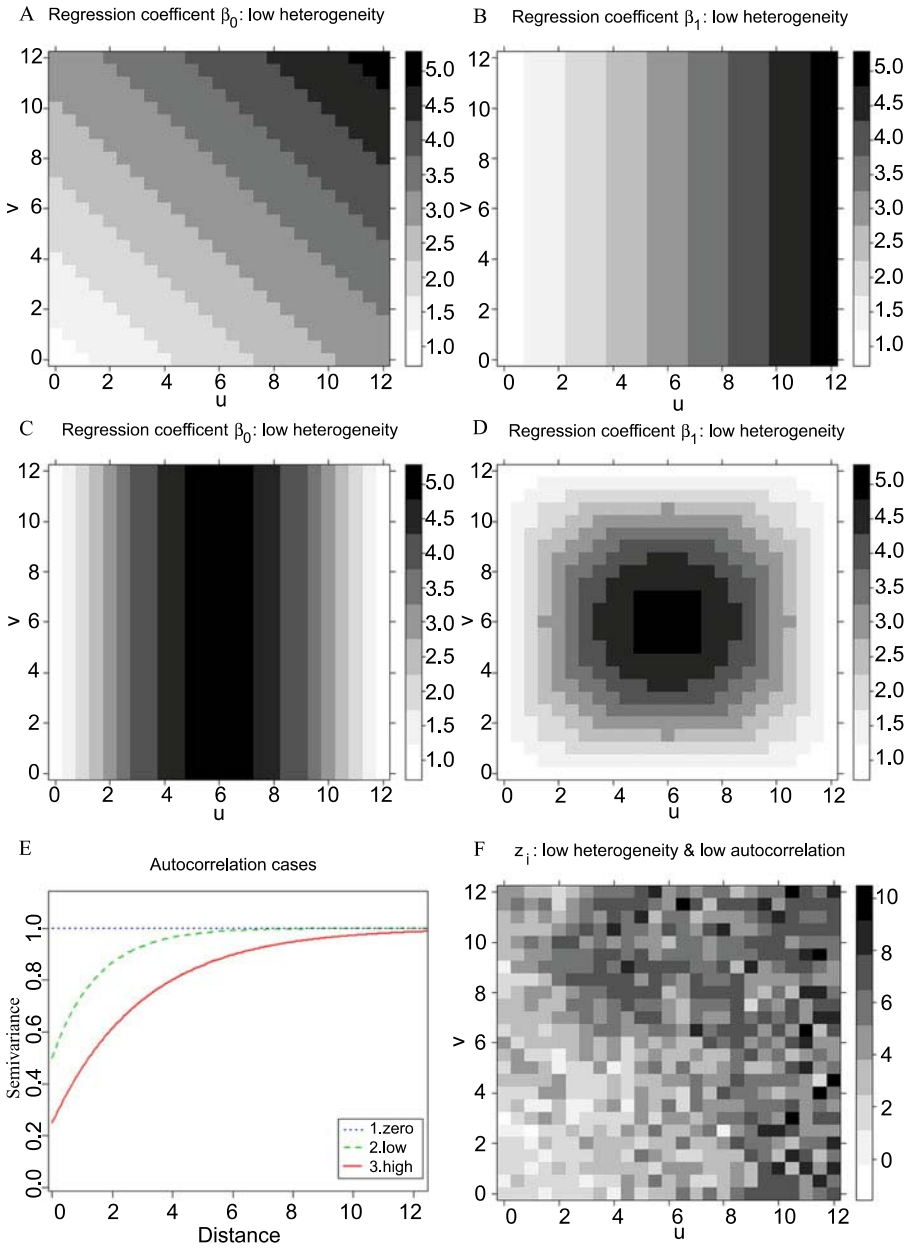

Fig. 1 Simulated data: True regression coefficient surfaces for heterogeneity cases 2 (**A** and **B**) and 3 (**C** and **D**). Variograms for the three autocorrelation cases (**E**) and an example output of the z_i data for simulated class 5 (**F**)

- Case 1, zero autocorrelation: independent draws from a normal distribution $N(0, (\text{var}(m(\mathbf{x}))/2)^{0.5})$.

Table 1 Simulated classes and heterogeneity/autocorrelation indicators

Mean/residual process	Zero heterogeneity	Low heterogeneity	High heterogeneity
Simulated classes:			
Zero autocorrelation	1	4	7
Low autocorrelation	2	5	8
High autocorrelation	3	6	9
Indicators of heterogeneity:			
Zero autocorrelation	58.29	10.56	3.96
Low autocorrelation	4.84	3.32	2.44
High autocorrelation	2.64	2.51	1.96
Indicators of autocorrelation:			
Zero autocorrelation	0.000	0.350	0.331
Low autocorrelation	0.053	0.420	0.411
High autocorrelation	0.068	0.453	0.428

Relatively small heterogeneity indicator values suggest simulated data have strong levels of heterogeneity
Relatively large autocorrelation indicator values suggest simulated data have strong levels of autocorrelation

- Case 2, low autocorrelation: data generated according to an isotropic exponential variogram model with $c_0 = (1/2)(\text{var}(m(\mathbf{x}))/2)$, $c_1 = (1/2)(\text{var}(m(\mathbf{x}))/2)$ and $a = 1.5$ distance units. This gives a nugget effect (defined as $c_0/(c_0 + c_1)$) of 0.50.
- Case 3, high autocorrelation: data generated according to an isotropic exponential variogram model with $c_0 = (1/4)(\text{var}(m(\mathbf{x}))/2)$, $c_1 = (3/4)(\text{var}(m(\mathbf{x}))/2)$ and $a = 3$ distance units. This gives a nugget effect of 0.25.

Observe how the nugget effects are generated, where further investigations could assess stronger and weaker nugget effects. For clarity, each of the three autocorrelation cases is depicted in Fig. 1E.

Once all the residual data sets are generated, the corresponding dependent data z_i can be found, which results in nine different data sets, each with different heterogeneity/autocorrelation properties (or classes, see Table 1) from the same simulation run. Thus each simulation run consists of (u_i, v_i, z_{si}, y_i) data, where $s = 1, 2, \ldots, 9$. An example z_i surface is given in Fig. 1F. Observe that the nine data sets that are generated do not have the same mean or variance, but the simulation algorithm is such that these parameters are broadly similar. Here mean values can range from 4 to 5 and variance values can range from around 2 to 7.

4 Model Calibration and Validation

For each simulated data set, model calibration and validation is conducted using a set-aside procedure where a random sample of 310 observations is used for model calibration and the remaining 315 observations are used for model validation (i.e. an approximate 50:50 split). The same data division is used for data within each

simulation run, but not for data from different simulation runs. The (five) resultant data configurations were not considered an important model discriminating factor in subsequent assessments.

4.1 Model Calibration

For practical and objectivity reasons, all model parameters are found in a fully-automated fashion. Sensible heuristics are used for the starting parameters of the REML or WLS variogram fits and pure nugget models can be specified where necessary. Monitoring of the progress of model calibration is also done at strategic points. For example, the RMSE bandwidth/neighbourhood functions are checked for instances of (a) multiple minima and (b) exact local collinearity where models fail (although such events are unlikely with our simulated data—see Sect. 6.3). The automated algorithm performed reliably, where the ability to check model calibration was considered valuable. However this user-input entailed that data from each of the nine classes were only simulated five times (i.e. five simulation runs, giving $9 \times 5 = 45$ simulated data sets in total).

4.2 Model Validation: Prediction Accuracy

Due to the nature of the simulation algorithm, a relative prediction accuracy diagnostic is calculated, as this enables a comparison of model performance across all nine simulated data classes. In particular, a relative RMSE diagnostic is reported, which is defined as $\text{relRMSE} = \text{RMSE}_{\text{Model}}/\text{RMSE}_{\text{Mean}}$. The further this ratio falls below unity, the greater the improvement in precision over using the calibration mean as the predictor of the corresponding validation data set. RMSE values are calculated in a similar way to that defined in Sect. 2.1, but now accord to a set-aside procedure. For each model and simulation data class, relRMSE results are reported as an average of the individual relRMSE values found from each simulation run.

4.3 Model Validation: Prediction Uncertainty Accuracy

For all models, an assumption of multivariate normality is adopted. This entails that the prediction $\hat{z}(\mathbf{x})$ and the prediction standard error (PSE) $\sigma(\mathbf{x})$ can be taken as the two defining moments of a normal distribution at location $\mathbf{x}$, which in turn enables the calculation of a prediction confidence interval (PCI) whose accuracy can be assessed using coverage probabilities. For example, if symmetric 95% PCIs were calculated at each validation location (i.e. using $\hat{z}(\mathbf{x}) \pm 1.96\sigma(\mathbf{x})$) then a correct modelling of local uncertainty would entail that there is a 0.95 probability that the actual value $z(\mathbf{x})$ falls within the interval. In other words, 95% of the symmetric 95% PCIs should contain the actual value. Furthermore, if a coverage probability is found for a range of symmetric PCIs (say from a 1% to a 99% PCI in increments of 1%) and the results are plotted against the probability interval p, then an accuracy plot is found (Deutsch 1997; Goovaerts 2001). Sample accuracy plots are given for GWR-EXP-CV, OK, UK-LN, and GWRK-EXP-AIC in Fig. 2, where the results for MLR are shown in all plots to provide context. For this particular simulation run and class, the performance

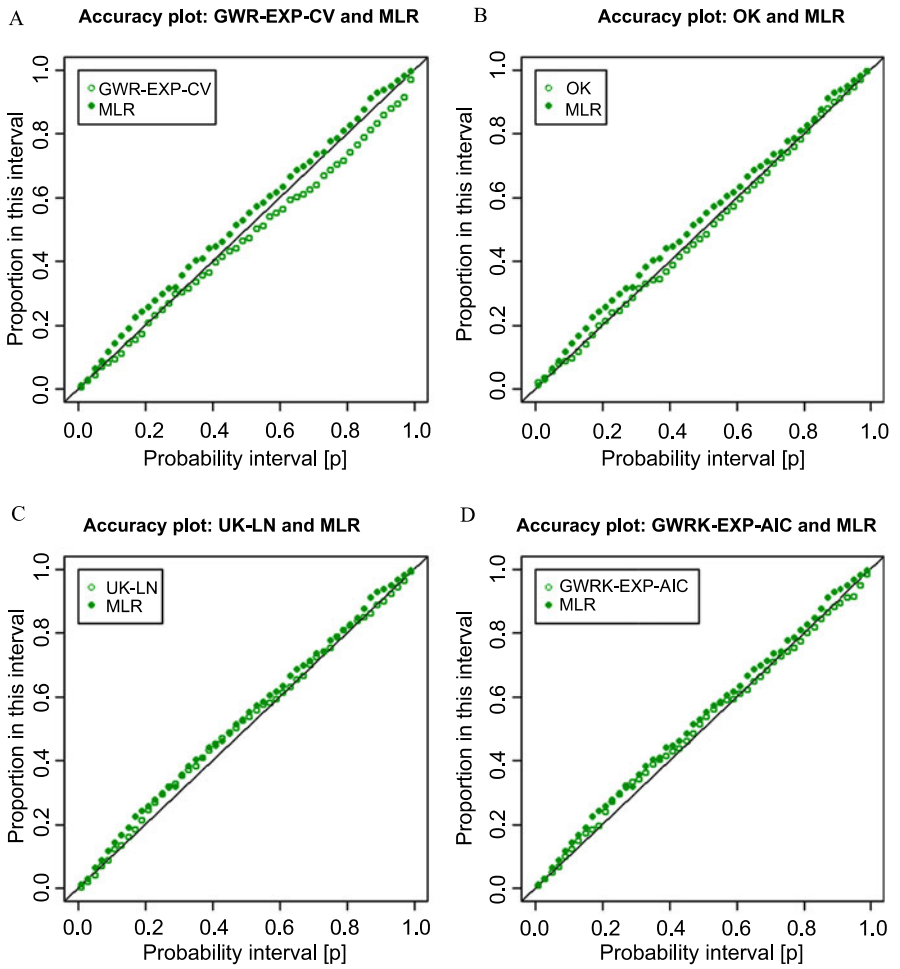

Fig. 2 Accuracy plots for (**A**) GWR-EXP-CV, (**B**) OK, (**C**) UK-LN, and (**D**) GWRK-EXP-AIC using simulation data from run three for simulated class 5 (low heterogeneity and low autocorrelation cases)

of GWR-EXP-CV is relatively weak, as its accuracy plot falls furthest from the 45° line.

Following Deutsch (1997), a complementary goodness statistic is calculated, which provides value to a model's accuracy plot. This G-statistic can be defined as $G = 1 - \int_0^1 [3a(p) - 2][\bar{\xi}(p) - p]\,\mathrm{d}p$, where $\bar{\xi}$ is the fraction of actual values falling in the PCI, and a value of 1 is sought. The indicator function $a(p)$ is defined as $a(p) = \begin{cases} 1 & \text{if } \bar{\xi}(p) \geq p \\ 0 & \text{otherwise} \end{cases}$, which entails that twice the importance is given to deviations when $\bar{\xi}(p) < p$. Thus, models with accuracy plots that tend to fall above the 45° line are preferred to models that have accuracy plots that tend to fall below the 45° line. For each model and simulation data class, G-statistic results are reported as an average of the individual G-statistic values found from each simulation run.

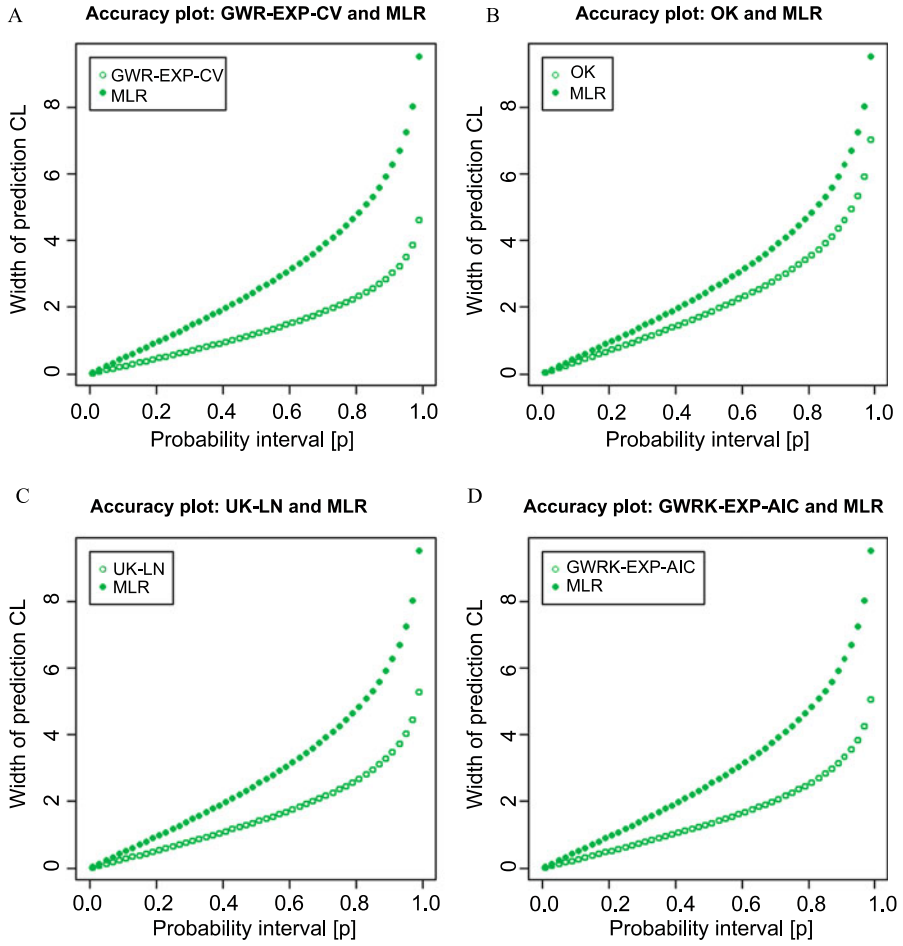

Fig. 3 PCI width plots for (**A**) GWR-EXP-CV, (**B**) OK, (**C**) UK-LN, and (**D**) GWRK-EXP-AIC using simulation data from run three for simulated class 5 (low heterogeneity and low autocorrelation cases)

For cases where two models provide similar G-statistics, one model can be preferred if its PCI widths that contain the actual value are smaller, as this reduces uncertainty about the actual value falling within the PCI. Therefore, the average width of the PCIs that include the actual values is calculated for each p and plotted in Fig. 3 for the same models used in Fig. 2. Clearly, MLR provides very poor PCI's in this respect. To cater to this important aspect of a PCI, an average PCI width for all p (AW) is calculated for each model and simulation data class, and again reported as an average of the five individual AW values found.

4.4 Model Validation: Ranked Summaries

A general view on model performance across all simulation data classes is found by presenting the average relRMSE, G-statistic, and AW results in a ranked form.

Here the mean rank of a model is defined as $\mathrm{MRK}_m = (1/9) \sum_{s=1}^{9} r_{ms}$, where r_{ms} is the performance rank (via its average relRMSE, G-statistic, or AW value) of the mth model with simulation data class s. The SD of the ranks can be found using $\mathrm{SDRK}_m = [(1/(9-1)) \sum_{s=1}^{9} (r_{ms} - \mathrm{MRK}_m)^2]^{0.5}$. Models that perform well should have a small mean rank and a small SD of ranks, with respect to each assessment statistic. A large SD of ranks indicates that a model tends to perform either best or worst.

4.5 Ball-park Indicators of Spatial Heterogeneity and Spatial Autocorrelation

To aid in the interpretation of the results, ball-park indicators of spatial heterogeneity and spatial autocorrelation that are actually present in each (full) simulated z_i data set is provided (Table 1). An indicator of heterogeneity is taken as the bandwidth for a GWR-BX-CV model averaged over the five simulation runs. An indicator of autocorrelation is taken as a Moran's I statistic (specified with an IDW scheme), averaged over the five simulation runs. In the former case, (relatively) small values would suggest simulated data with strong heterogeneity. In the latter case, (relatively) large values would suggest simulated data with strong autocorrelation. From the behaviour of the indicator data, it is clear that each component of the simulation has a direct effect on the other. Here increasing heterogeneity in the mean component increases autocorrelation in the simulated z_i data; similarly, increasing autocorrelation in the residual component increases heterogeneity in the simulated z_i data. Thus weak or strong levels of heterogeneity tend to coincide with weak or strong levels of autocorrelation, respectively. This effect is akin to the usual identification problem of the decomposed model, where it not possible to isolate first- from second-order effects. However, the problem is further complicated as local non-linear trends are defined (see Sect. 6.4). Consequently, when interpreting our study results, it is better to view the indicator data rather than the labels of the simulation class for context. Furthermore, we should not expect our simulated experiment to distinguish between data situations when a GWR predictor can be preferred to a kriging predictor (and vice versa), as both will tend to perform relatively well or poorly within the same simulated data class.

5 Results

Details of the performance of the twelve models in nine different simulated data classes (reflecting variations in both spatial heterogeneity and spatial autocorrelation) are given in Tables 2–4 for the three different goodness-of-fit measures described in Sect. 4, where each measure (relRMSE, G-statistic, and AW) is averaged over five simulations. The results are presented in the same format as that given in Table 1, for each of the twelve models. For example, the relRMSE values (Table 2) for OK for simulated classes 1, 5, and 9 are 1.001, 0.681, and 0.561, respectively. To aid the interpretation of the results, the three best performing models within each model-type are highlighted in bold, and the three best performing models within each simulated class are filled in. For example, the three best performing MLR models with respect

Table 2 Validation results for relRMSE

OK			MLR			UK-GN			UK-LN		
1.001	0.757	0.694	**0.589**	0.884	0.953	0.592	0.611	0.623	0.592	0.594	0.620
0.983	0.681	**0.599**	**0.567**	0.904	0.941	0.508	0.531	0.490	0.508	0.508	0.481
0.975	**0.642**	**0.561**	**0.537**	0.885	0.938	**0.392**	**0.449**	**0.439**	**0.393**	**0.408**	**0.415**
GWR-BX-CV			**GWR-EXP-CV**			**GWR-BX-AIC**			**GWR-EXP-AIC**		
0.592	0.604	0.660	0.590	0.599	0.655	0.590	0.604	0.665	0.590	0.603	0.687
0.543	0.534	0.544	0.524	0.530	0.517	0.547	0.537	0.540	**0.529**	0.539	0.576
0.440	**0.451**	**0.485**	**0.427**	**0.427**	**0.467**	**0.440**	**0.449**	**0.484**	**0.440**	**0.464**	0.547
GWRK-BX-CV			**GWRK-EXP-CV**			**GWRK-BX-AIC**			**GWRK-EXP-AIC**		
0.596	0.603	0.656	0.593	0.598	0.654	0.594	0.601	0.651	0.593	0.601	0.644
0.517	0.528	0.518	0.517	0.525	0.516	0.516	0.523	0.519	0.514	0.523	0.501
0.426	**0.432**	**0.460**	**0.422**	**0.425**	**0.465**	**0.421**	**0.432**	**0.464**	**0.410**	**0.421**	**0.446**

The three best performing models within each model-type are in bold
The three best performing models within each simulated class are highlighted
Each entry of each model's 3 × 3 sub-table directly relates to those given in Table 1

Table 3 Validation results for G-statistic

OK			MLR			UK-GN			UK-LN		
0.776	0.966	**0.977**	**0.979**	0.962	0.967	**0.976**	0.972	**0.979**	**0.976**	**0.972**	**0.981**
0.937	0.956	**0.984**	**0.972**	0.931	0.955	0.961	0.948	0.974	0.961	0.965	0.968
0.912	0.968	**0.978**	**0.974**	0.953	0.943	0.970	**0.975**	0.974	0.970	0.969	0.969
GWR-BX-CV			**GWR-EXP-CV**			**GWR-BX-AIC**			**GWR-EXP-AIC**		
0.976	0.972	0.966	**0.978**	**0.953**	0.853	**0.977**	0.970	0.967	**0.978**	**0.960**	0.944
0.969	0.951	**0.973**	**0.936**	0.885	0.864	0.965	0.951	**0.975**	0.953	0.925	**0.966**
0.968	0.966	**0.980**	0.872	0.852	0.763	0.972	0.965	**0.981**	0.956	0.919	0.945
GWRK-BX-CV			**GWRK-EXP-CV**			**GWRK-BX-AIC**			**GWRK-EXP-AIC**		
0.964	0.958	0.944	**0.963**	**0.935**	0.842	0.964	0.951	0.949	0.963	0.938	0.926
0.974	0.949	**0.973**	**0.932**	0.887	0.852	**0.975**	0.963	**0.974**	0.954	0.930	**0.977**
0.966	0.965	**0.974**	0.870	0.846	0.751	0.967	0.964	**0.975**	**0.966**	0.948	**0.980**

The three best performing models within each model-type are in bold
The three best performing models within each simulated class are highlighted
Each entry of each model's 3 × 3 sub-table directly relates to those given in Table 1

to relRMSE (Table 2) are those for simulated classes 1, 2, and 3; the three best performing models within simulated class 1 with respect to relRMSE (Table 2) are MLR, GWR-EXP-CV, and GWR-EXP-AIC.

Using relRMSE as a performance indicator, three models seem to do consistently well under varying conditions of heterogeneity and autocorrelation: UK-LN, GWRK-EXP-AIC, and UK-GN. As expected, MLR only performs reasonably well under conditions of very weak autocorrelation. Conversely, OK performs very poorly in such cases, but its performance improves with increasing heterogeneity and autocorrelation. As expected, all models (aside from the aspatial MLR model) perform better with increased levels of heterogeneity and autocorrelation. Using the G-statistic as a performance indicator is less useful in discriminating between models. Few if any consistent conclusions can be drawn from the results and there is little to choose between the models under a large variety of conditions. AW values can only be viewed as a performance indicator in conjunction with the G-statistic (and similarly, a model's G-statistic is given value when reported with AW). Thus, although GWR-EXP-CV, GWRK-EXP-CV, and GWRK-EXP-AIC appear to perform well in this

Table 4 Validation results for AW

OK			MLR			UK-GN			UK-LN		
3.659	2.425	2.724	**2.153**	2.880	3.674	2.146	1.964	2.401	2.152	1.977	2.416
3.443	**2.139**	2.406	**2.002**	2.734	3.731	1.758	**1.602**	2.071	1.759	**1.613**	2.086
3.315	**1.893**	**2.023**	**1.856**	2.616	3.329	**1.358**	**1.328**	1.603	**1.360**	**1.338**	1.625
GWR-BX-CV			**GWR-EXP-CV**			**GWR-BX-AIC**			**GWR-EXP-AIC**		
2.161	1.928	2.498	2.149	1.795	1.992	2.156	1.923	2.501	2.150	1.852	2.417
1.907	**1.652**	2.201	1.692	1.408	1.650	1.909	**1.656**	2.190	1.775	**1.546**	2.150
1.489	**1.347**	1.734	**1.193**	**1.030**	**1.132**	**1.496**	**1.336**	1.723	**1.408**	**1.267**	1.756
GWRK-BX-CV			**GWRK-EXP-CV**			**GWRK-BX-AIC**			**GWRK-EXP-AIC**		
2.047	1.844	2.390	2.035	1.734	1.953	2.041	1.821	2.379	2.036	1.760	2.204
1.837	**1.627**	2.175	1.659	1.399	1.615	1.835	**1.637**	2.174	1.736	**1.510**	2.015
1.472	**1.327**	1.715	**1.177**	**1.012**	**1.108**	**1.468**	**1.510**	1.691	**1.366**	**1.214**	1.623

The three best performing models within each model-type are in bold
The three best performing models within each simulated class are highlighted
Each entry of each model's 3 × 3 sub-table directly relates to those given in Table 1

Table 5 Ranked performances

Model:	1—relRMSE		2—G-statistic		3—AW value	
	MRK_m	SDRK_m	MRK_m	SDRK_m	MRK_m	SDRK_m
OK	11.33	0.50	5.67	3.94	11.33	0.50
MLR	10.56	3.61	5.67	3.57	11.44	1.01
UK-GN	4.11	3.26	4.44	2.24	5.11	1.96
UK-LN	1.89	1.96	4.33	2.96	6.56	2.07
GWR-BX-CV	8.22	0.97	4.22	1.56	9.33	0.87
GWR-EXP-CV	5.22	1.86	9.89	3.06	2.56	1.33
GWR-BX-AIC	8.33	1.80	3.78	1.48	9.11	1.17
GWR-EXP-AIC	8.56	2.35	8.33	2.18	6.33	1.87
GWRK-BX-CV	6.11	2.03	6.67	1.94	6.44	1.51
GWRK-EXP-CV	4.78	1.79	11.67	0.71	1.00	0.00
GWRK-BX-AIC	5.33	2.12	5.67	3.00	5.67	1.73
GWRK-EXP-AIC	3.56	2.13	7.67	3.46	3.11	0.93

respect, they each have relatively poor G-statistic results. An example of the use of AW value would be to distinguish between GWR-BX-CV, UK-LN, and UK-GN for simulated data class 4 where their G-statistics are the same (at 0.972). Here GWR-BX-CV has the lowest AW value (at 1.928) and therefore provides the better estimates of prediction uncertainty for this particular example. Instances where models perform particularly well in this dual assessment are with UK-GN and GWRK-EXP-AIC for simulated data classes 3 and 8, respectively.

Ranked summaries (Table 5) provide an alternative guide to each model's relative performance. Thus from a prediction accuracy perspective (i.e. relRMSE), the overall best model is UK-LN followed by GWRK-EXP-AIC and UK-GN. From a prediction uncertainty accuracy perspective and using the G-statistic only, GWR-BX-AIC, GWR-BX-CV, and UK-LN are the three best performing models. However, when

Table 6 Estimated model parameters: average bandwidths/neighbourhoods (%)

OK			MLR			UK-GN			UK-LN		
–	–	–	–	–	–	–	–	–	79.03	22.26	23.55
–	–	–	–	–	–	–	–	–	94.52	24.84	24.84
–	–	–	–	–	–	–	–	–	68.71	23.55	13.23
GWR-BX-CV			GWR-EXP-CV			GWR-BX-AIC			GWR-EXP-AIC		
54.73	15.68	5.23	87.48	1.80	0.36	71.86	16.19	6.61	86.71	2.94	1.75
26.63	5.69	5.87	1.47	0.93	0.33	29.45	8.75	5.48	3.18	1.90	1.51
6.02	5.64	4.50	0.63	0.44	0.26	9.93	5.48	4.25	1.87	1.68	1.47

Each entry of each model's 3×3 sub-table directly relates to those given in Table 1

a model's AW ranked performance is considered jointly with its G-statistic ranked performance then UK-GN, GWR-BX-AIC, and GWRK-BX-AIC are viewed as the three best performing models. MLR, OK, and GWR-EXP-AIC are viewed as the three worst performing models in this respect.

The relatively strong prediction accuracy performances of UK-LN and GWRK-EXP-AIC should be expected as both models directly account for both heterogeneity and autocorrelation. Thus, the performance of these models is in part a consequence of the simulation approach adopted. The performance of UK-GN is commendable and indicates that substantial first-order effects (via different levels of relationship heterogeneity) can be successfully modelled as second-order effects instead. Furthermore, and unlike UK-LN and GWRK-EXP-AIC, UK-GN does not suffer from some form of approximation with respect to its estimates of prediction uncertainty (and hence its strong performance in this respect).

GWR models are almost always outperformed by their corresponding GWRK models. Performance difference is greatest between GWR-AIC and corresponding GWRK-AIC models (which reflects their parameterisation as discussed below). The relatively strong performance of the GWRK-AIC models is expected as their GWR-AIC mean components will more strongly tend to MLR fits than corresponding GWR-CV fits will, which in turn entails that GWRK-AIC models will more strongly tend to (the strongly-performing) UK-GN (or RK) models than corresponding GWRK-CV models will. It is difficult to gauge whether the performance difference between the best GWRK model (GWRK-EXP-AIC) and the best GWR model (taken as GWR-EXP-CV) is sufficiently large to counteract the differences in model complexity. For GWRK to be preferred, it must not only show a distinct advantage over GWR, but also the simpler UK-GN model. Basic models are not only preferable, in the interest of model parsimony, but are also preferable with respect to statistical inference (i.e. it is difficult to interpret residual variogram bias statistically in GWRK, whereas for UK-GN it is viable). Pardo-Igúzquiza and Dowd (1998) discuss such issues when preferring UK-GN to more complicated or highly parameterised models.

For completeness, average bandwidths or neighbourhoods for the GWR or UK-LN models are given in Table 6 for the nine simulated data classes. Similarly, aver-

Table 7 Estimated model parameters: average nugget effects

OK			MLR			UK-GN			UK-LN		
0.792	0.006	0.002	–	–	–	0.769	0.005	0.002	0.769	0.005	0.002
0.387	0.005	0.011	–	–	–	0.471	0.005	0.007	0.471	0.005	0.007
0.763	0.008	0.032	–	–	–	0.252	0.005	0.007	0.252	0.005	0.007
GWRK-BX-CV			GWRK-EXP-CV			GWRK-BX-AIC			GWRK-EXP-AIC		
0.870	0.862	0.870	0.872	0.898	0.926	0.869	0.855	0.849	0.872	0.867	0.767
0.771	0.857	0.773	0.834	0.860	0.899	0.731	0.797	0.799	0.791	0.792	0.658
0.827	0.790	0.806	0.877	0.892	0.899	0.806	0.798	0.826	0.743	0.695	0.583

Each entry of each model's 3×3 sub-table directly relates to those given in Table 1

age nugget effects for the OK, UK and GWRK models are given in Table 7 to give a partial guide to modelled levels of autocorrelation. Again, these parameter estimates reflect an intertwined heterogeneity/autocorrelation effect in the data (cf. the indicator data in Table 1). Here bandwidths/neighbourhoods tend to shrink and nugget effects (for the standard OK/UK models) tend to strengthen, as levels of heterogeneity/autocorrelation increase.

Observe that the GWRK models are consistently modelled with a high nugget effect, which was expected. Observe also that GWRK-CV models tend to have higher nugget effects than GWRK-AIC models. This reflects the usual stronger fit of a GWR-CV mean component over a corresponding GWR-AIC mean component (i.e. view GWR relRMSE and bandwidth results in Tables 2 and 6, respectively), leaving a weaker (and likely more biased) autocorrelation structure in GWRK-CV residual data than that found in corresponding GWRK-AIC residual data. Furthermore, such differences tend to be pronounced when an exponential rather than a box-car kernel is specified. Observe that there is no evidence to suggest that any of the GWR-CV models have over-fitted the simulated data. If such events do occur, then the corresponding GWR-AIC models often provide reasonable alternatives.

6 Discussion

6.1 Modelling Decisions

Throughout this study, the decisions taken when parameterising a given model are taken to be both reasonable and consistent. Some effort was made to look at different model forms where, for example, GWR was specified with two different kernel functions and with different optimised bandwidths. However, a more complete comparison of model form could potentially have looked at: GWR with further kernel options (e.g. bi-square, etc.); GWRK with different residual variogram estimators; OK/UK/GWRK with different variogram models (e.g. Matérn, etc.); OK/UK/GWRK with different variogram model fitting techniques; etc. Investigating all such model forms has the potential to change the results of this study, but clearly such an in-depth

comparison would soon become rather tedious and unwieldy. Automating modelling decisions is possible, but some severe computational difficulties may result. As such, the results of this study have to be viewed in context of the modelling decisions made, as well as the particular properties of the simulated data generated. In our opinion, it is felt that the results are reasonable where the most important modelling decisions were adequately investigated.

6.2 Estimated PSEs

The majority of this study's models have been shown to provide reasonably accurate PSE estimates in a global (and statistical) sense (Tables 3 and 4). However, it can also be useful to assess the value of a model's PSE estimates spatially and locally. In this respect, maps of absolute (actual) prediction errors and estimated PSEs are given in Figs. 4 and 5, for the same four models used in Figs. 2 and 3. Clearly, there is no relationship between actual errors and those estimated for each of the chosen models; this result can be more clearly observed from the corresponding scatterplots of Fig. 6. All that is evident is that kriging provides PSEs that reflect sample configuration (i.e. PSEs are higher in areas of sparsely sampled data). These results are transferable to all other study models and are not unexpected.

For kriging, this problem has been a topic of much debate (e.g. Journel 1986; Goovaerts 2001; Heuvelink and Pebesma 2002), and a common approach to address it is to specify some non-stationary model that allows the variance and/or variogram to vary locally. Commonly, such models improve local prediction uncertainty accuracy but have little or no effect on prediction accuracy. Models include kriging with locally varying sills (Isaaks and Srivastava 1989), Box-Cox kriging (Kitanidis and Shen 1996), kriging with an interpolation variance (Yamamoto 2000) and MWK. GWR can be similarly adapted and our current research is exploring this issue with the spatially heteroskedastic models of Fotheringham et al. (2002, pp. 80–82) and of Páez et al. (2002), where estimates of the residual variance are allowed to vary across space.

6.3 Empirical Research and Model Calibration Issues with UK-LN

For the combined heterogenic and autocorrelation properties of our simulated data sets, results suggest that UK-LN should be chosen ahead of GWRK, which itself should be chosen ahead of UK-GN (where this order reflects a higher weighting is given to prediction accuracy rather than prediction uncertainty accuracy). Thus, when it comes to transferring these results to real data sets that have similar spatial properties, UK-LN should be the preferred predictor. However, our own empirical research has routinely confirmed a common problem with the calibration of UK-LN (and by association GWR-BX and GWRK-BX) models where an optimal neighbourhood cannot be found when covariates are not particularly continuous (e.g. Deutsch and Journel 1998, p. 71) and/or are locally collinear (i.e. models fail across a range of neighbourhoods due to matrix instability). Crucially, GWR/GWRK models specified with some continuous distance-decay kernel are usually able to circumvent such problems and as a result, are more flexible than UK-LN when modelling non-stationary relationships. For instances where UK-LN models could not

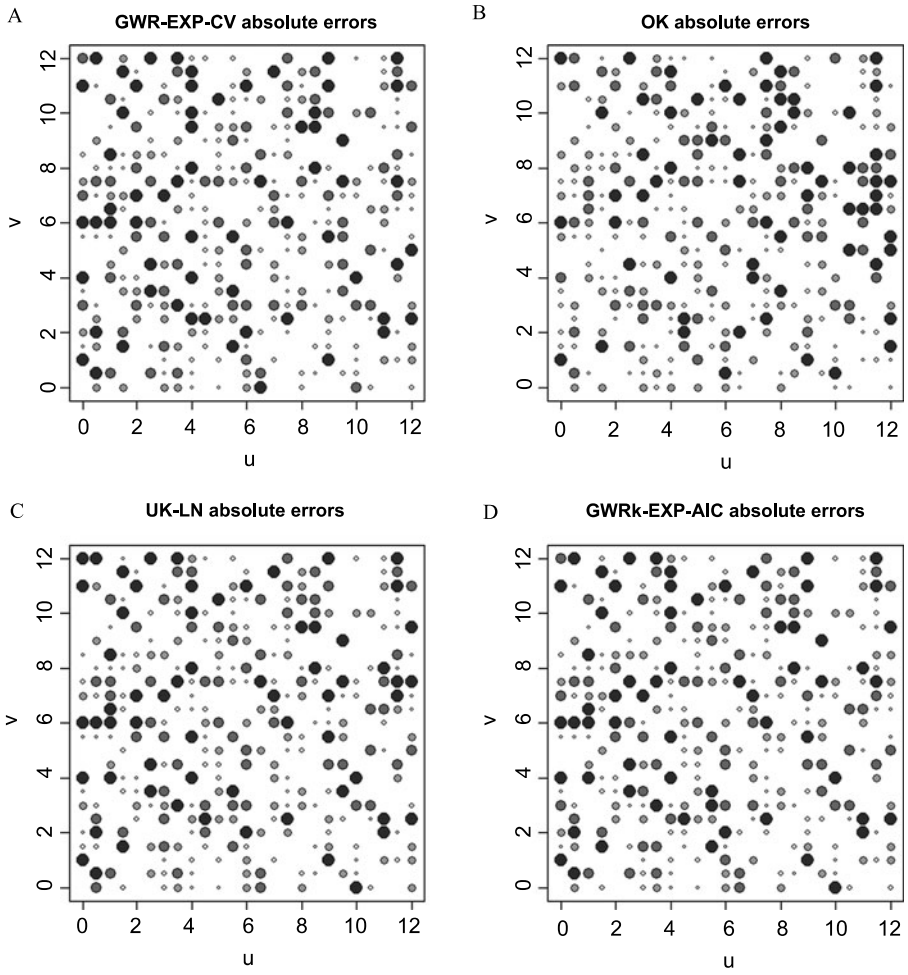

Fig. 4 Absolute prediction error maps for (**A**) GWR-EXP-CV, (**B**) OK, (**C**) UK-LN, and (**D**) GWRK-EXP-AIC using simulation data from run three for simulated class 5 (low heterogeneity and low autocorrelation cases). Large dark-coloured filled circles correspond to large errors whereas small light-coloured filled circles correspond to small errors

be calibrated, GWRK-EXP (and GWR-EXP) models have performed favourably against UK-GN models, especially from a prediction accuracy perspective. For instances where problems of local collinearity cannot be adequately addressed with basic GWR constructions, the sophisticated ridge or lasso GWR models of Wheeler (2007, 2009) can be used instead, as these are specifically designed to counter such problems.

6.4 Limitations of the GWRK Model

Finally, it is important to stress that there is much uncertainty regarding the GWRK model. The proposed GWRK variances are (ad hoc) rough approximations (see (9))

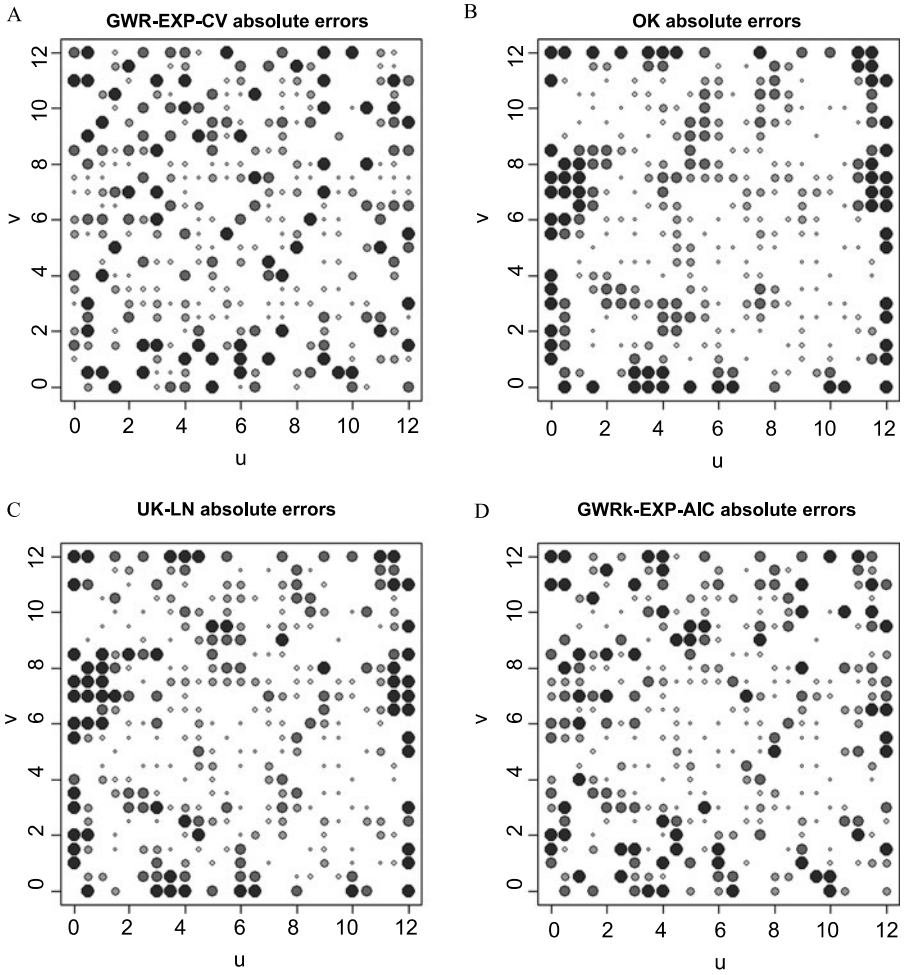

Fig. 5 Estimated prediction standard error maps for (**A**) GWR-EXP-CV, (**B**) OK, (**C**) UK-LN, and (**D**) GWRK-EXP-AIC using simulation data from run three for simulated class 5 (low heterogeneity and low autocorrelation cases). Large dark-coloured filled circles correspond to large estimated errors whereas small light-coloured filled circles correspond to small estimated errors

and bias in the residual variogram was not investigated; problems that both relate to a difficult identification problem that is inherent in this model. As a consequence, the GWRK model is not ready for any inferential analysis, as there are concerns about it its stability, correlation and statistical consistency. These concerns may be addressed by further research, possibly borrowing ideas from the formulation of UK as a random effects model as in Kammann and Wand (2003). Thus for now, GWRK should be viewed as a useful prediction model only.

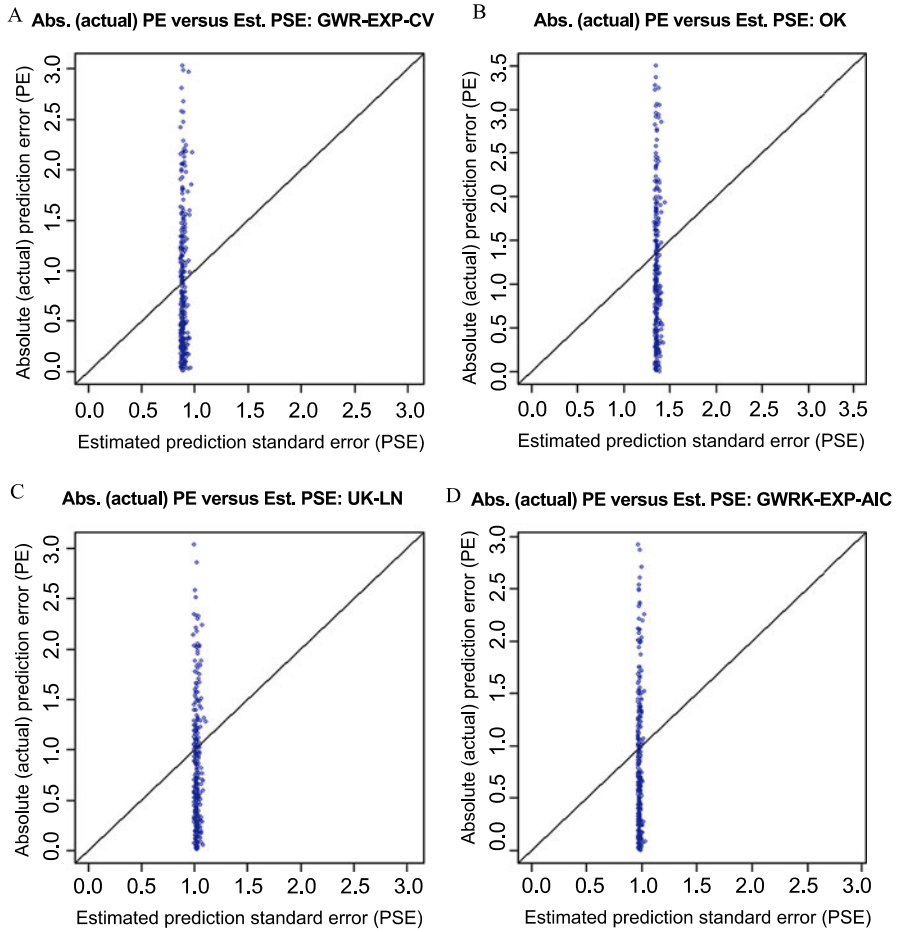

Fig. 6 Actual versus estimated prediction error plots for (**A**) GWR-EXP-CV, (**B**) OK, (**C**) UK-LN, and (**D**) GWRK-EXP-AIC using simulation data from run three for simulated class 5 (low heterogeneity and low autocorrelation cases)

7 Conclusions

In this study, we have compared the prediction performances of MLR, GWR, OK, UK, and GWRK (GWR used as a mean component of kriging) models using simulated data based on different levels of spatial heterogeneity (with respect to data relationships) and different levels of spatial autocorrelation. The best performing model was found to be a UK model specified with local neighbourhoods. This model was followed by a GWRK model and then by a UK model specified with a global neighbourhood. All UK, GWRK, and GWR models were shown to out-perform the naïve MLR and OK models. For the three best performing models, all three could account for spatial autocorrelation, but only the first two could additionally account for spatially varying relationships. Although our simulation experiment does not prove a result, these results can provide a useful guide to empirical research. In this respect,

it is noted that UK models specified with local neighbourhoods (our preferred choice) can sometimes suffer from calibration difficulties and if so, a GWRK (or sometimes, a more pragmatic GWR) model can provide a worthy alternative when predicting with non-stationary relationships.

Further comparative work could consider using simulated data with different levels (and types) of heterogeneity and autocorrelation to that used here. Investigations into the effects of sample size and configuration on model performance should also be considered, as would the simulation of data based on more than one covariate. Future research could also incorporate non-stationary relationship models that have not been investigated in this study, such as the local cokriging (CoK) models of Pereira et al. (2002); the UK- and CoK-based MWK models of Haas (1996); the Bayesian spatially varying coefficient models of Gelfand et al. (2003); and the Bayesian local CoK models of Gelfand et al. (2004).

Acknowledgements Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support. We also thank the anonymous referees whose thoughtful comments and insights helped to improve the paper.

References

Atkinson PM (2001) Geographical information science: geocomputation and nonstationarity. Prog Phys Geogr 25:111–122

Bitter C, Mulligan GF, Dall'erba S (2007) Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. J Geogr Syst 9:7–27

Chilès JP, Delfiner P (1999) Geostatistics—modelling spatial uncertainty. Wiley, New York

Cressie N (1986) Kriging nonstationary data. J Am Stat Assoc 81:625–634

Deutsch CV (1997) Direct assessment of local accuracy and precision. In: Baafi EY, Scofield NA (eds) Geostatistics Wollongong '96. Kluwer Academic, Dordrecht, pp 115–125

Deutsch CV, Journel AG (1998) GSLIB geostatistical software library and user's guide. Oxford University Press, New York

Farber S, Páez A (2007) A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. J Geogr Syst 9:371–396

Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression—the analysis of spatially varying relationships. Wiley, Chichester

Gambolati G, Volpi G (1979) A conceptual deterministic analysis of the kriging technique in hydrology. Water Resour Res 15:625–629

Gao X, Asami Y, Chung C (2006) An empirical evaluation of spatial regression models. Comput Geosci 32:1040–1051

Gelfand AE, Kim HJ, Sirmans CJ, Banerjee S (2003) Spatial modeling with spatially varying coefficient processes. J Am Stat Assoc 98:387–396

Gelfand AE, Schmidt AM, Banerjee S, Sirmans CJ (2004) Nonstationary multivariate process modeling through spatially varying coregionalisation. Test 13:266–312

Genton MC, Furrer R (1998) Analysis of rainfall data by robust spatial statistics using S+SPATIALSTATS. J Geogr Inf Decis Anal 2:126–136

Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. Geoderma 103:3–26

Haas TC (1990) Lognormal and moving window methods of estimating acid deposition. J Am Stat Assoc 85:950–963

Haas TC (1996) Multivariate spatial prediction in the presence of non-linear trend and covariance nonstationarity. Environmetrics 7:145–165

Hengl T, Heuvelink GBM, Rossiter DG (2007) About regression kriging: from equations to case studies. Comput Geosci 33:1301–1315

Heuvelink GBM, Pebesma EJ (2002) Is the ordinary kriging variance a proper measure of interpolation error? In: Hunter G, Lowell K (eds) The fifth international symposium on spatial accuracy assessment in natural resources and environmental sciences. RMIT University, Melbourne, pp 179–186
Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York
Journel AG (1986) Geostatistics: models and tools for the earth sciences. Math Geol 18:119–140
Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic Press, London
Kammann EE, Wand MP (2003) Geoadditive Models. J R Stat Soc C, Appl Stat 52(1):1–18
Kanevski M, Maignan M (2004) Analysis and modeling of spatial environmental data. Dekker, New York
Kitanidis PK, Shen KF (1996) Geostatistical interpolation of chemical concentration. Adv Water Resour 19:369–378
Leung Y, Mei C, Zhang W (2000) Statistical tests for spatial nonstationarity based on the geographically weighted regression model. Environ Plann 32:9–32
Lloyd CD (2010) Nonstationary models for exploring and mapping monthly precipitation in the United Kingdom. Int J Climatol 30:390–405
Matheron G (1963) Principles of geostatistics. Econ Geol 58:1246–1266
Matheron G (1969) Le krigeage universal. Centre de Morphologie, Fontainebleau
Matheron G (1989) Estimating and choosing: an essay on probability in practice. Springer, Berlin
Neuman SP, Jacobson EA (1984) Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. Math Geol 16:499–521
Páez A, Uchida T, Miyamoto K (2002) A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity. Environ Plann 34:733–754
Páez A, Long F, Farber S (2008) Moving window approaches for hedonic price estimation: an empirical comparison of modelling techniques. Urban Stud 45:1565–1581
Pardo-Igúzquiza E, Dowd PA (1998) The second-order stationary universal kriging model revisited. Math Geol 30:347–378
Pereira MJ, Soares A, Rosario L (2002) Characterization of forest resources with satellite spot images by using local models of co-regionalization. In: Kleingeld WJ, Krige DG (eds) Geostatistics 2000. Geostatistical Association of Southern Africa, Cape Town
Rivoirard J (2002) On the structural link between variables in kriging with external drift. Math Geol 34:797–808
Schabenberger O, Gotway C (2005) Statistical methods for spatial data analysis. Chapman & Hall, London
Wang N, Mei C, Yan X (2008) Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. Environ Plann 40:986–1005
Wheeler D (2007) Diagnostic tools and a remedial method for collinearity in geographically weighted regression. Environ Plann 39:2464–2481
Wheeler D (2009) Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. Environ Plann 41(3):722–742
Yamamoto JK (2000) An alternative measure of the reliability of ordinary kriging estimates. Math Geol 32:489–509
Zhang X, Eijkeren JC, Heemink AW (1995) On the weighted least-squares method for fitting a semivariogram model. Comput Geosci 21:605–608
Zhang L, Gove JH, Heath LS (2005) Spatial residual analysis of six modeling techniques. Ecol Model 186:154–177
Zimmerman DL, Pavik C, Ruggles A, Armstrong MP (1999) An experimental comparison of ordinary and universal kriging and inverse distance weighting. Math Geol 31:375–390