Computers, Environment and Urban Systems
26 (2002) 501–524

Computers, Environment and Urban Systems

www.elsevier.com/locate/compenvurbsys

Geographically weighted summary statistics — a framework for localised exploratory data analysis

C. Brunsdon *, A.S. Fotheringham, M. Charlton

Spatial Analysis Research Group, Department of Geography, University of Newcastle-upon-Tyne, Newcastle-Upon-Tyne, NE1 7RU, UK

Abstract

Geographical kernel weighting is proposed as a method for deriving local summary statistics from geographically weighted point data. These local statistics are then used to visualise geographical variation in the statistical distribution of variables of interest. Univariate and bivariate summary statistics are considered, for both moment-based and order-based approaches. Several aspects of visualisation are considered. Finally, an example based on house price data is presented. # 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Geographical kernel weighting; Localised exploratory data analysis

1. Introduction

One of the first subjects covered in elementary statistics courses is that of summary statistics, or descriptive statistics. These are usually introduced as a means of data reduction (Ehrenberg, 1982). For example, a mean and a standard deviation can give information about the spread and location of a database containing a very large number of measurements. This general approach can be a very useful tool — indeed some do not feel the need to progress to more sophisticated forms of statistical analysis. However, for geographers these statistics have a major shortcoming. They may be thought of as whole-map statistics (Openshaw, 1991). Rather than giving information about spatial variation within the study region, the entire data set (the 'whole map') is summarised as a single entity. For the spatial analyst, classical data

* Corresponding author.
E-mail address: chris.brunsdon@ncl.ac.uk (C. Brunsdon).

PII: S0198-9715(01)00009-6

reduction perhaps provides too much reduction — summarising a data set with a map may be more helpful than with a pair of numbers.

This problem is perhaps best understood with a practical example. Consider the county of Tyne and Wear in the UK, shown in Fig. 1. In Fig. 2, the agreed sale prices of 1067 houses in and around the county of Tyne and Wear, UK in 1991[1] are shown in the form of a stem-and-leaf diagram. A map of the study area is provided in Fig. 1. The sale price has a mean value of £43,472 and a standard deviation of £22,642. However, this does not imply that one could select any neighbourhood in

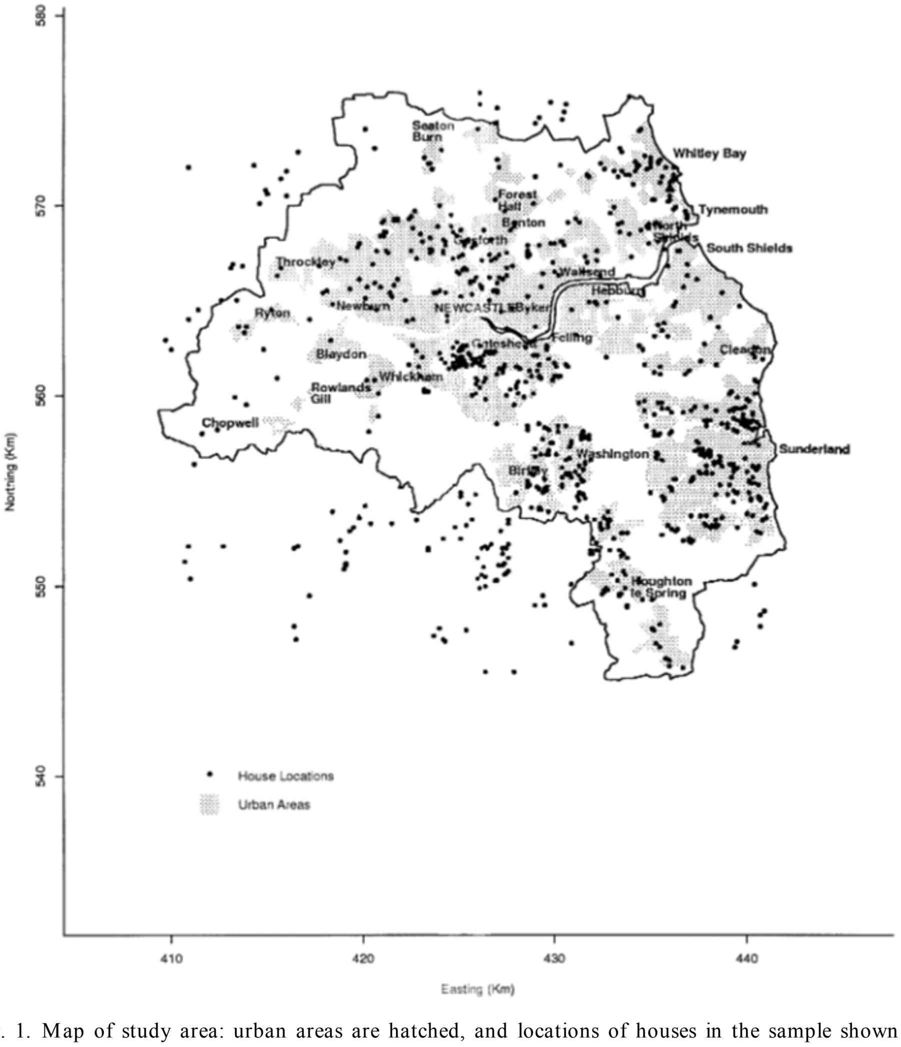
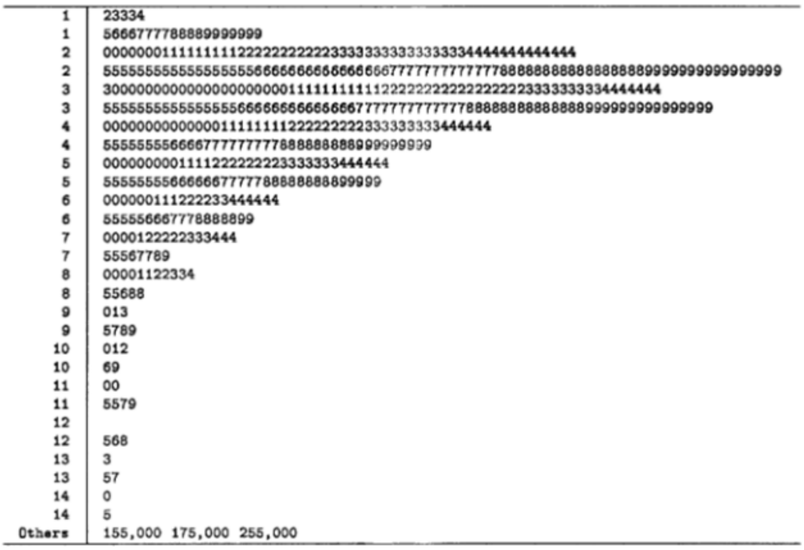


Fig. 1. Map of study area: urban areas are hatched, and locations of houses in the sample shown as points. Note some houses are outside of the study area to reduce the problems of edge effects.

[1] We gratefully acknowledge the Nationwide Building Society for allowing us to use this data.



```
     1 | 23334
     1 | 5666777788889999999
     2 | 0000000111111112222222222223333333333333333444444444444
     2 | 555555555555555555566666666666666667777777777777788888888888888888889999999999999999
     3 | 30000000000000000000000111111111112222222222222222223333333333444444
     3 | 55555555555555555566666666666666777777777777788888888888888888899999999999999999
     4 | 000000000000001111111122222222222333333333444444
     4 | 55555555666677777777778888888888999999999
     5 | 00000000011112222222222333333334444444
     5 | 55555555666666777778888888888899999
     6 | 0000001112222233444444
     6 | 5555566677788888899
     7 | 0000122222333444
     7 | 55567789
     8 | 00001122334
     8 | 55688
     9 | 013
     9 | 5789
    10 | 012
    10 | 69
    11 | 00
    11 | 5579
    12 |
    12 | 568
    13 | 3
    13 | 57
    14 | 0
    14 | 5
Others | 155,000 175,000 255,000
```

Fig. 2. Stem-and-leaf plot of house sale price data from Tyne and Wear. Numbers in stem are multiples of 10,000, leaves are multiples of 1000. Each number represents two observations.

Tyne and Wear and expect a typical house price of £43,472, or a 'spread' of house prices characterised by a standard deviation of £22,642. Local variations in the housing market are very important to consider.

In this paper we outline a method which allows a wide variety of summary statistics to be localised — so that for any point (u,v) in the study area a summary for a small area around that point can be obtained. The approach has parallels with that of geographically weighted regression (GWR; Brunsdon, Fotheringham, & Charlton, 1996; Fotheringham, Charlton, & Brunsdon, 1998). As with GWR, the approach is based on weighted statistics, with each observation in the data set being weighted in terms of its proximity to (u,v).

However, it is not just the mean and standard deviation that can be treated in this way. A glance at Fig. 2 suggests that the distribution of house sale prices is not symmetrical — there is a long upper tail. The skewness of the distribution is a descriptive statistic which conveys this information at a whole-map level, but again it might be helpful to investigate localised skewness. Maps of this would help to investigate whether the skewness seen at the global level is reflected locally, or whether the upper tail seen globally is due to a few estates having an unskewed collection of expensive housing. Also of interest are order-based descriptive statistics, such as the median and inter-quartile range. These tend to be more robust to outliers. Here, we outline an approach where localised versions of these may also be derived.

The advantages of this approach over simple choropleth maps as a data exploration tool are twofold. Firstly, the well-known difficulties associated with data aggregation to a pre-specified system of areal units is avoided. Openshaw collectively terms these difficulties the Modifiable Areal Unit Problem (MAUP) and provides a comprehensive discussion of their scope and nature (Openshaw, 1984). Secondly, it provides a direct way of viewing changes in the degree of variability of more general 'shape' of data distributions over a geographical space — such effects can only be inferred from a choropleth map, and indeed the perception of these may well be subject to the MAUP.

This paper sets out the approach in two stages. Firstly we give an overview of the basic method, and then outline a general theoretical perspective which may be used to derive localised versions of a large set of summary statistics. Secondly we consider a set of global descriptive statistics which might be appropriate candidates for localisation, and then, using results from Section 2, a set of definitions of localised statistics are drawn up. Having considered practical aspects of computation and visualization, the paper concludes with a more detailed consideration of the house sale price data introduced earlier.

2. Localised descriptive statistics

As mentioned above, the key to the localisation method described here is geographical weighting. This has been previously used in a regression context (Brunsdon et al., 1996; Fotheringham, Charlton, & Brunsdon, 1997) and extended to spatial regression models (Brunsdon, Fotheringham, & Charlton, 1998). Initially, we describe the basic methodology for adapting global statistics to local ones via geographical weighting, and then we consider the process in more theoretical depth. The latter provides a conceptual framework for deriving localised versions of a very general range of statistics.

In a sense, the work reported here is akin to the broader subject of spatial interpolation, and it is worth discussing this linkage. Methods for interpolating spatial data are widely used in geographical research. These methods are usually available in GIS software and are often used, for example, in interpolating values at the mesh points of a regular grid from irregularly spaced sample data. Remote sensing software provides another rich source of functions for processing data that are collected on a regular lattice (for example high-pass filters for noise removal). It might therefore be argued that the statistics proposed here offer little that is new. However the research reported here forms part of a general methodology that can be applied to a wide range of descriptive statistics. In order to put this assertion into context it is desirable to review briefly some of the more common methods for spatial interpolation.

A number of useful summaries of what methods are available for spatial interpolation have been compiled (Lam, 1983) and (Burrough, 1986). One classification of methods is to consider those that are useful for irregularly spaced sample data (for example, spot heights obtained from field survey), and conversely, those that are useful for data collected on a regular lattice (for example a remotely sensed image). Another classification of interpolators is based on whether the interpolated surface passes exactly through the sample data, or whether the results are approximate (assuming some model of global trend). Yet another view of interpolation arises by considering the sampling points to be the result of some stochastic process. Finally we may consider whether methods are appropriate for categorical or continuous attributes.

One frequently used set of techniques assumes that that the spatial variation in the attribute arises as the result of some moving average process. This requires that the analyst specifies some sample size for the computation of the moving average and some form of distance weighting function so that near sample points have a greater weight in the resulting interpolation than far sample points. The inputs to such interpolation procedures are the values of an attribute at a set of irregularly spaced locations, and the output is usually, but not always, the values of an attribute interpolated at a set of regularly spaced locations. With the data for a raster image, the attribute value may be required at the mesh points of a grid that is rotated, or skewed from those at which the sensors in the satellite obtained the data. Such geometric correction employs similar methods to those outlined above depending on the type of data in the image. Whilst proximal methods may be reasonable for an already classified image, distance based methods are more suited to preservation of the original data values.

Another group of interpolators useful with regularly spaced sample locations (for example raster data in a GIS) is provided by Tomlin's Map Algebra moving window functions (Tomlin, 1990). The focalmedian function, for example, will return the median value of those in a window of user defined size. However, the output from these functions is intended to be located only at the mesh points of the regular grid.

The localised descriptive statistics proposed here make use of some of these concepts, notably the distance weighted interpolations and moving window methods, but relax the need to pre-specify a window size. They are also useful with both regularly and irregularly spaced locations for input and output. In addition to extending these ideas, we intend to show how localised descriptive statistics may be derived by considering the probability density of the attribute of interest conditioned on the geographical location via the technique of Kernel density estimation (Silverman, 1986).

3. Concepts of localised descriptive statistics

In principal, the calibration of a statistical model is localised to a point (u,v) by weighting each observation in the data set according to its proximity to (u,v). For instance, if the ith data point is situated at point (u_i,v_i), then w_i, the weight applied to the ith point, could be defined by:

$$w_i = \exp(-d_i^2/h^2) \qquad (1)$$

where d_i is the Euclidean distance between point i and (u,v). Points close to (u,v) are highly weighted, and this weighting reduces as d_i increases, tending to zero as d_i becomes very large. The parameter h — the bandwidth — controls the rate at which this fall-*off* in weighting occurs. Note that the weighting scheme changes as (u,v) moves. Thus, as (u,v) scans an entire study area, calibration of the model 'focuses' on points in the locality of (u,v). Any model parameter estimated in this way becomes a continuous function[2] of (u,v), and can thus be represented graphically by a surface or contour map. The bandwidth — whose dimension is that of a distance — effectively determines the 'tightness' of the focus. In the weighting function in Eq. (1), observations at a distance 2h from (u,v) give a w_i of about 0.02, which is relatively low. Since the weighting varies with (u,v), we refer to the technique as geographical weighting.

Here we apply geographical weighting to a variety of descriptive statistics. For example the sample mean may be replaced by a locally weighted mean:

$$\bar{x}(u,v) = \frac{\sum x_i w_i(u,v)}{\sum w_i(u,v)} \qquad (2)$$

where the w_i's are determined by Eq. (1). The (u,v) notation after $\bar{x}$ and w_i serves to indicate that these quantities vary as (u,v) changes. For brevity, this will be omitted for the w_i's in the rest of this paper — it will be assumed that all w_i's in this paper depend on (u,v) unless otherwise stated. Equations such as Eq. (2) can also be simplified if the w_i's are re-scaled to sum to one, which they may be without loss of generality. In this case we define the new w_i to be $\frac{w_i}{\sum w_i}$ (in terms of original w_i's). Again we will assume that the w_i's have been scaled to sum to unity throughout the paper, unless otherwise stated. Thus, Eq. (2) may be re-written as:

$$\bar{x}(u,v) = \sum x_i w_i \qquad (3)$$

In fact, this is simply an interpolation formula (Ripley, 1981). However, the geographical weighting approach may be extended beyond this. For example, a geographically weighted standard deviation may be defined as:

$$s_x(u,v) = \sqrt{\sum (x_i - \bar{x}(u,v))^2 w_i} \qquad (4)$$

Note the use of $\bar{x}$(u,v) in this definition. Locally weighted variation around the localised mean is of interest here, not locally weighted variation around the global mean.

With Eqs. (3) and (4) we have a basic toolkit for exploring geographical variation in statistical distributions. Evaluating this summary statistic for all points in the study area yields a surface — which can be mapped. In Figs. 3 and 4, contour maps of localised versions of the mean and standard deviation are shown for the house price data, with h= 3 km. The mean value is lowest (below £30,000) in a central area surrounding Newcastle and Gateshead, in the southernmost tip of the county around Hetton-le-Hole and also towards the west. High mean values, in excess of £60,000, are seen in the north west corner of the county. Similar comments can be made in terms of the localised standard deviation — generally higher values (in

[2] Provided the weighting function is continuous.

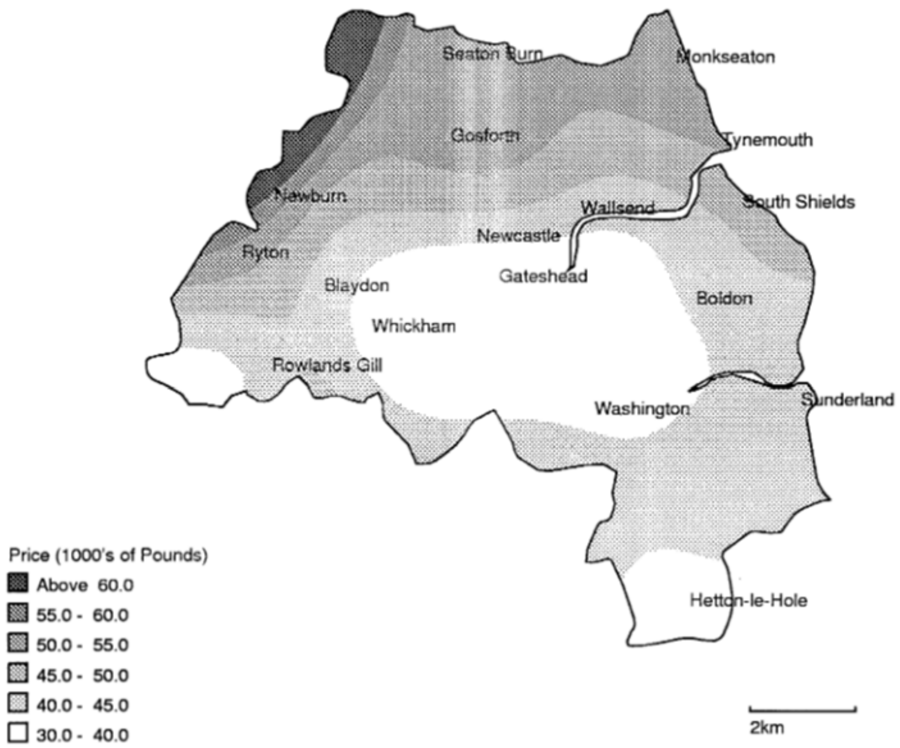


Fig. 3. Localised mean for house price data.

excess of £26,000) occur where the local mean is highest, and low values where the local mean is lowest. However, there are exceptions to this rule — for example a ‘dip’ between Gosforth and Tynemouth is evident in the localised standard deviation, but not in the localised mean. Note that, although the study area is restricted to the county itself, some data is taken from areas just beyond the county borders — this is to avoid the problems of ‘edge effects’ — where sparse data close to the edge of the study area may cause spurious patterns to appear. When considering only global descriptive statistics the features of the Tyneside housing market shown in Figs. 3 and 4 would remain hidden.

However, as suggested in Section 1, these two examples are not the only possibilities. The obvious extension is to consider a geographically weighted skewness, defined in terms of the third moment about the (local) mean:

$$b_x(u,v) = \sqrt[3]{\frac{(x_i - \bar{x}(u,v))^3 w_i}{[s_x(u,v)]^3}} \qquad (5)$$

Note that, as with the mean, it is the localised version of the standard deviation that is used in the formula.

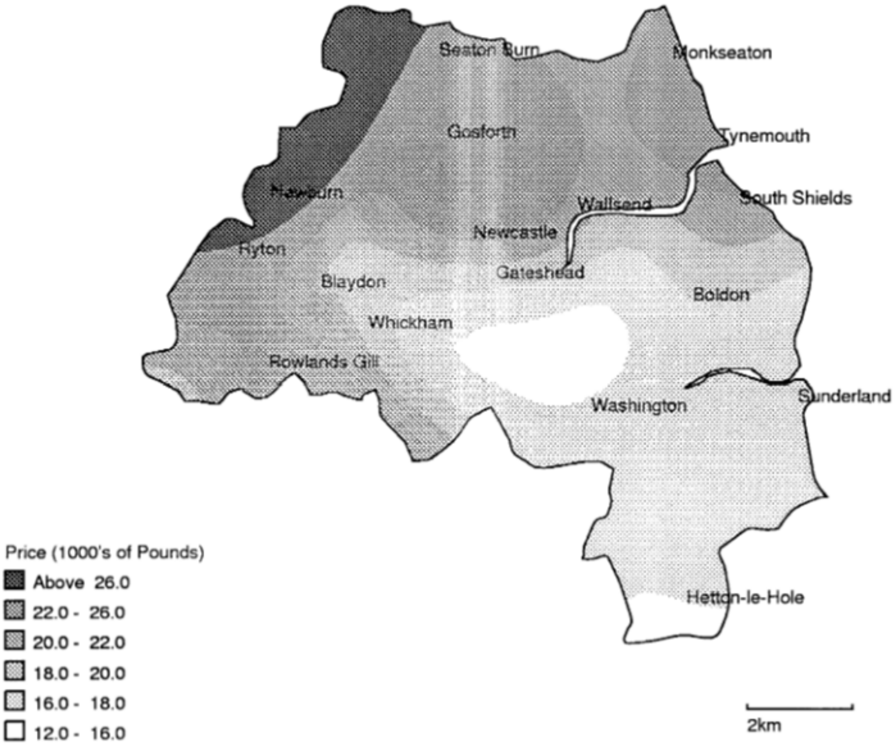


Fig. 4. Localised standard deviation for house price data.

Thus, using geographical weighting allows us to localise a number of moment-based[3] summary statistics. In each case, the localised statistics are generalisations of the global ones — if n is the sample size, and each w_i is equal to n^{-1} in any of expressions (3)–(5), then these expressions would coincide with the unweighted definitions of the statistics, regardless of (u,v). Thus, the global statistic is just a special case of the geographically weighted statistic. In the following subsection we demonstrate how generalisations of this form may be applied to a much broader range of summary statistics.

4. A theoretical perspective

Here, we give a theoretical interpretation of the simple geographical weighting approach outlined above. In order to do this, we must first consider the probability distribution of the data triplet (u,v,x). This is a three-dimensional quantity giving the location of a data point, together with its x-value. Suppose the probability density function is f(u,v,x). Here, we consider the location and the x-value to be random

[3] i.e. Based on expressions of the form $\sum (x_i - c)^m w_i$.

quantities — for example, if x is a house sale price then f(u,v,x) is the probability density that some house sale takes place at location (u,v) and is sold for x. This function contains information about the relative frequencies of house sales at different (u,v)-locations, and also about the likely value of house price sales at different (u,v)-locations. However, when considering local descriptive statistics, we are only interested in the latter of these two factors. To focus attention on this, we consider the conditional distribution of x given (u,v). This is written as f(x|u,v), and using a standard result we have:

$$f(x|u,v) = \frac{f(u,v,x)}{\int_x f(u,v,x)dx} \tag{6}$$

This is just the joint probability function with u and v treated as known values, rescaled so that $\int f(x|u,v)dx = 1$. Thus, f(x|u,v) is essentially a univariate distribution in x. By comparing these probability densities for different (u,v)'s, we can see how the local probability density for x varies geographically. Hence, we define the term localised density of x to be the density f(x|u,v).

Now we return to summary statistics. Since localised densities are essentially univariate densities, theoretical localised summary statistics may be defined by applying the global summary statistic definition to localised densities. Thus, a localised mean can be defined as:

$$E(x|u,v) = \int_x x f(x|u,v)dx. \tag{7}$$

Table 1 shows several definitions of summary statistics for a general univariate density f(x), and a general discrete probability distribution Pr(x). Note that the

Table 1
Typical descriptive statistics for the univariate probability density function f(x)

Statistic name	Definition		Notation
	Continuous	Discrete	
Mean	$\int xf(x)dx$	$\sum x\Pr(x)$	E(x)
Standard deviation	$\int (x-E(x))^2 f(x)dx$	$\sum (x-E(x))^2\Pr(x)$	SD(x)
Skewness	$\frac{\int (x-E(x))^3 f(x)dx}{SD(x)^{1.5}}$	$\frac{\sum (x-E(x))^3 \Pr(x)}{SD(x)^{1.5}}$	Sk(x)
p-Quantile	Solution for q of $\int_{-\infty}^{q} f(x)dx = p$	Minimum solution for q of PR(x< q)= p	$Q_p(x)$
Median	$Q_{0.5}(x)$		Med(x)
Inter-quartile range	$Q_{0.75}(x)-Q_{0.25}(x)$		IQR(x)
Quantile imbalance	$\frac{2Med(x)-(Q_{0.75}(x)+Q_{0.25})}{IQR(x)}$		QI(x)

notation for these distribution-based measures is different from those based on the sample data considered in Section 3. As it has been argued, all of these summary statistics may be localised, by substituting f(x|u,v) for f(x) in this table. Thus, at least for theoretical distributions, we have a way to define localised versions not only for the moment-based mean, standard deviation and skewness statistics considered earlier, but also for quantile-based statistics such as the median and the interquartile range. Better still, we have a way of generating a localised version of any statistic defined in terms of a univariate density function.

This characteristic sets this method apart from parametric model based techniques, such as Kriging (Krige, 1966; Matheron, 1973). Kriging allows smooth prediction surfaces to be computed which are similar to the locally weighted mean. Some variants on the technique also allow for local changes in the geostatistical characteristics of the model (Goovaerts, 1997). However these approaches assume that the distribution of x is Gaussian. Thus, it is implicitly assumed that local skewness is always zero, and that the local mean is always equal to the local median. It is possible to transform x but this would then assume that the distribution of the transformed x applied everywhere. In some situations, the distribution may vary from place to place so that in some places a transformation might be appropriate, but not in other places. Here, we relax the Gaussian assumption by using a non-parametric model — we simply state that the local distribution of x is f(x|u,v), without restricting f to any functional form — or indeed to the same function form for all u and v. In the situation where there is prior knowledge that the distribution is Gaussian (with or without transformation), Kriging is likely to give a more efficient estimate. However, in an exploratory context where no such prior knowledge exits, the locally weighted methods set out here are more appropriate tools.

In practice we normally have a set of observed data, not a theoretical distribution. How can we proceed in this case? One approach is to generate an estimate of the underlying distribution f(u,v,x) based on the data, and then compute localised summary statistics by first deriving an estimate of f(x|u,v) using Eq. (6), and then applying the appropriate distribution-based summary statistic formula. Since here the emphasis is based on an exploratory approach, we prefer not to adopt a parametric model for f(u,v,x). Instead, we use the non-parametric method of kernel density estimation (Brunsdon, 1995; Silverman, 1986; Wand & Jones, 1995) to estimate f directly from the data. We will not discuss the method in detail — consulting any of the above publications provides detailed discussions of the method. However, it is necessary to provide a basic outline. In the three dimensional case, a kernel density estimate $\hat{f}$(u,v,x) of the density f(u,v,x) is defined by:

$$\hat{f}(u,v,x) = \frac{1}{nh_u h_v h_x} \sum K\left(\frac{u-u_i}{h_u}, \frac{v-v_i}{h_v}, \frac{x-x_i}{h_x}\right) \qquad (8)$$

where K(u,v,x) is a probability density function with mean zero and variance one. Typically, it is unimodal, the mode is also located at zero and the function is symmetrical in that reversing the sign of any of u, v or x leaves the value of K(u,v,x) unaltered. A Gaussian distribution is a common choice, although this is not universal. Thus, the expression on the right hand side of Eq. (8) can be interpreted as the average of a set of 'humps' centered around each data point. The parameters h_u, h_v and h_x control the width the hump in the u, v and x directions. Like h in Section 3 these may be interpreted as bandwidths. Here, we assume that since u and v combine to form a two-dimensional space, the kernels should be isotropic in this space. This allows the results of the density estimation to be frame-independent in the sense that the coordinate system may be rotated or translated without altering the results of the analysis (Tobler, 1989). Thus, we set $h_u = h_v = h_{uv}$, giving:

$$\hat{f}(u,v,x) = \frac{1}{nh_{uv}^2 h_x}\sum K\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}};\frac{x-x_i}{h_x}\right) \quad (9)$$

Next, we assume that the kernel distribution K(u,v,x) may be factorised as $\alpha K_{uv}(u,v)K_x(x)$, where K_{uv} and K_x are, respectively, bivariate and univariate probability densities with mean zero and mode zero, and exhibit symmetry as defined above. α is a constant chosen to ensure that the K(u,v,x) integrates to one, and the variances of K_{uv} and K_x are chosen so that the variance of K is one. This is not too much of a restriction on K — many common choices of K fit this form, including the above mentioned Gaussian. If this assumption is made, we may write:

$$\hat{f}(u,v,x) = \frac{\alpha}{nh_{uv}^2 h_x}\sum K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}; K_x\frac{x-x_i}{h_x}\right) \quad (10)$$

Having reached this stage, we may substitute this expression for $\hat{f}(u,v,x)$ into Eq. (6), to obtain an estimate for the localised density $\hat{f}(x|u,v)$:

$$\hat{f}(x|u,v) = \frac{h_x^{-1}\sum K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}\right)K_x\left(\frac{x-x_i}{h_x}\right)}{\int_x h_x^{-1}\sum K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}\right)K_x\left(\frac{x-x_i}{h_x}\right)dx} \quad (11)$$

Note that the factor $\frac{\alpha}{nh_{uv}^2}$ appears in both numerator and denominator and is therefore canceled out. We do not cancel the factor h_x^{-1} for reasons which will become apparent below. Next, note that the factor not involving x in the denominator may be brought outside the integral, and that the order of the summation and the integral may be reversed so that the denominator may be written:

$$\sum K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}\right)\int_x h_x^{-1}K_x\left(\frac{x-x_i}{h_x}\right)dx \quad (12)$$

Now note that the integral in expression (12) is just the integral of a probability density function, and so is equal to one. Thus, expression (11) may be written as:

$$\hat{f}(x|u,v) = \frac{h_x^{-1}\sum K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}\right)K_x\left(\frac{x-x_i}{h_x}\right)}{\sum K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}\right)} \qquad (13)$$

This may be further simplified by setting:

$$w_i = \frac{K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}\right)}{\sum K_{uv}\left(\frac{u-u_i}{h_{uv}};\frac{v-v_i}{h_{uv}}\right)} \qquad (14)$$

to give:

$$\hat{f}(x|u,v) = h_x^{-1}\sum w_i K_x\left(\frac{x-x_i}{h_x}\right) \qquad (15)$$

Two important observations may now be made. Firstly, the w_i's defined above sum to one, and are based on a kernel function centred[4] on (u,v). They are therefore similar to the w_i's introduced in Section 3. In fact, if $K_x(x)$ is a Gaussian distribution, they are identical. Secondly, setting $w_i= n^{-1}$ in expression (15) gives the standard univariate kernel density estimator. Thus, expression expression (15) is a generalisation of a standard univariate kernel density in the same sense that the localised moment-based statistics in Section 3 are generalisations of their global versions. We will term the estimate in expression (15) the localised kernel density estimate. Like a standard kernel density estimate it is the average of distributions centered on the x-observations, but now the average is weighted in terms of the proximity of each observation to the point (u,v).

Tentatively, we now have a general method for computing localised summary statistics. Firstly, estimate the localised density with a localised kernel density estimate, and then substitute this into the expression for the statistic of interest. In fact, localised density estimates can be a useful exploratory tool in themselves. For example, using the house sales data from Fig. 2, setting h_{uv} at 3 km, and h_x at £12,000 and using Gaussian kernels, we compute localised kernel density estimates (Fig. 5) at the two locations whose grid references are (417.0,568.0) and (433.0,558.0). The former of these locations is in the Newburn area, and the latter is in Washington (Fig. 1). Inspecting the two curves in Fig. 5, it can be seen that the Newburn price distribution peaks at a higher value than that for Washington, and that it also has a long upper tail.

Next, we compute some summary statistics from $\hat{f}(x|u,v)$. The localised mean derived in this way can be found by taking means of both sides of expression (15):

[4] Until now K_{uv} was considered as being centered on (u_i,v_i). However, the symmetry of $K_{u,v}$ implies that (u,v) and (u_i,v_i) may be interchanged. Thus K_{uv} can also be thought of as centered on (u,v).

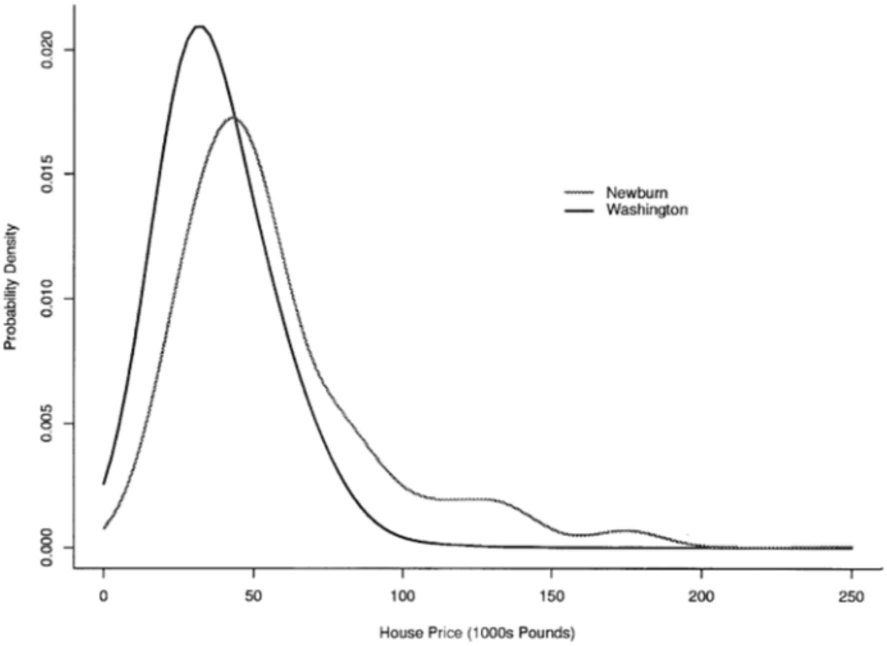


Fig. 5. Localised kernel density estimates for house sale price.

$$\hat{E}(x|u,v) = \sum w_i \int_x x h_x^{-1} K_x\left(\frac{x - x_i}{h_x}\right) dx. \qquad (16)$$

Noting that the integral on the right hand side is just the mean of the density $K_x(x)$ centered on x_i, which has the value x_i, we put:

$$\hat{E}(x|u,v) = \sum w_i x_i. \qquad (17)$$

Thus, the expression for the mean obtained in this way agrees with that obtained using the geographical weighting principal (Eq. (3)). Next we consider localised variance. This can be shown to be:

$$[\hat{s}d(x|u,v)]^2 = \sum w_i (x_i - \hat{E}(x|u,v))^2 + h_x^2. \qquad (18)$$

Note that this does not agree with the earlier definition given in Section 3 (Eq. (4)) unless $h_x = 0$. This implies an 'improper' kernel density estimate must be used to allow the two definitions to coincide. However, if one considers the limiting behaviour of kernel density estimates as $h_x \to 0$, it can be shown that the density estimate in expression (15) approaches a discrete distribution taking only the values $\{x_1 \ldots x_n\}$ with probabilities $\{w_1 \ldots w_n\}$. If this observation does not seem obvious, it may be helpful to view a practical example. In Fig. 6 the effect of allowing h_x to approach

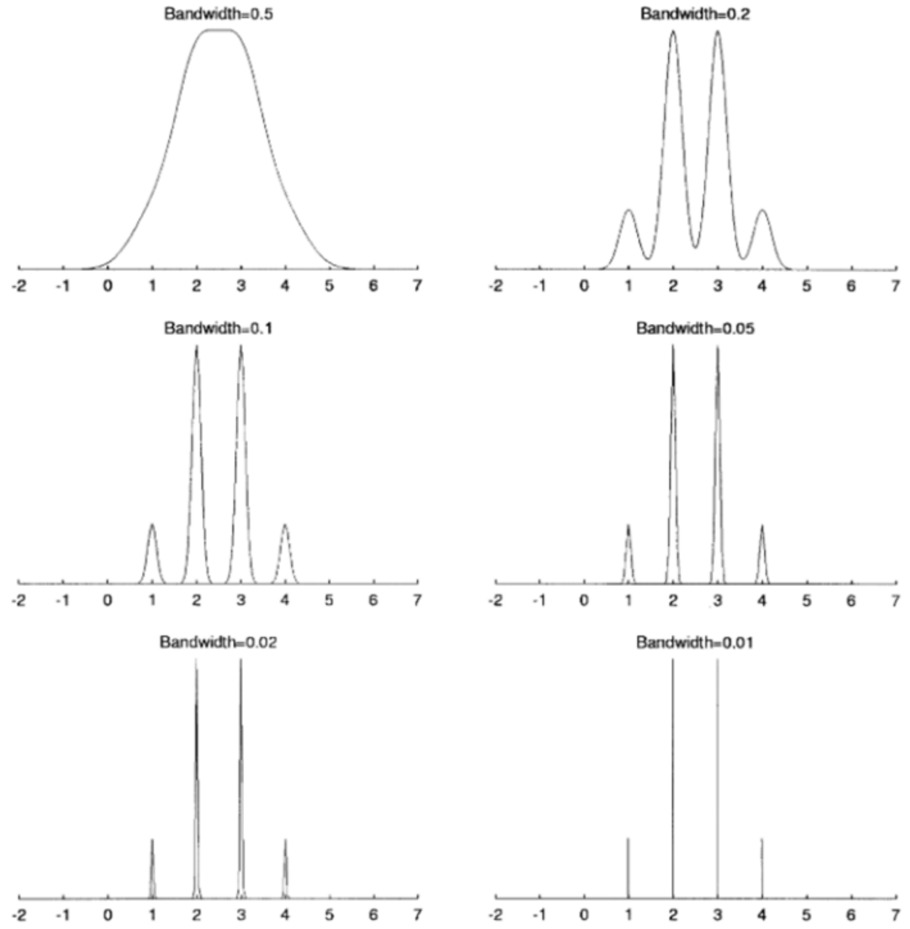


Fig. 6. *Effect* of letting bandwidth tend to zero in a localised kernel density estimate.

zero when the x-values are {1,2,3,4} and the corresponding w_i's are {0.1,0.4,0.4,0.1} is shown. For low values of h_x the graph approaches the form of four discrete 'mass points' of relative value given by the w_i's.

Thus, as a limiting case of a localised kernel density estimate, we have the n mass point distribution discussed above. Again, this is a localised distribution, as the w_i's all depend on (u,v). Observations close to (u,v) have high w_i's (that is, high masses), and observations further away from (u,v) have low w_i's (low masses). Essentially, these are geographically weighted discrete distributions, taking only the values $\{x_1 \ldots x_n\}$, with probabilities $\{w_1 \ldots w_n\}$ whose values reflect the proximity of the observations to the point (u,v). We term these localised mass point distributions (LMPDs). LMPDs for Newburn and Washington [using the same (u,v) values as in Fig. 5 and a bandwidth of 3 km with a Gaussian kernel] are shown in Fig. 7.

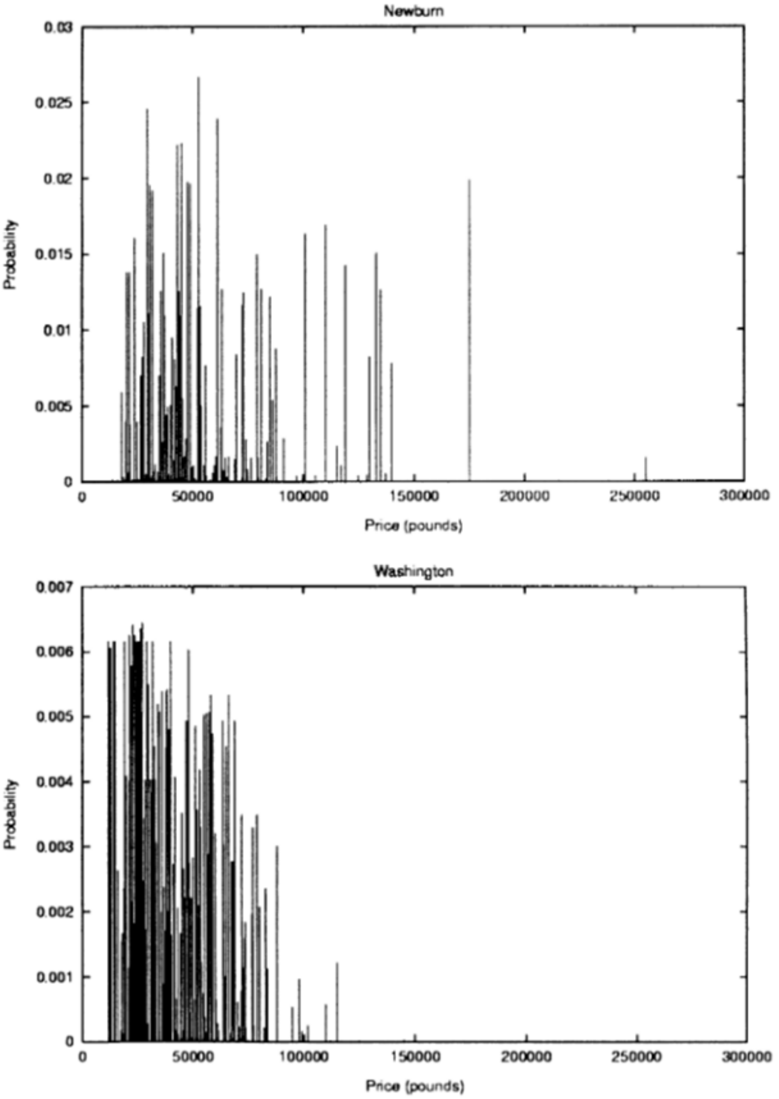


Fig. 7. Localised mass point distributions for Newburn and Washington.

These plots tell a similar story to the localised kernel density estimates considered earlier, with the Newburn distribution having a much longer 'tail' of expensive houses than Washington. The large spike at around £175,000 suggests that one house of this price was sold very close to the (u,v) sampling point used to represent Newburn here — there also appears to be a cluster of housing between about £90,000 and £130,000.

Returning to the theory, we have noted that LMPDs yield estimates of the localised mean and standard deviation which agree with those proposed in Section 3. In fact, it can be shown that this agreement holds for any moment-based statistics. Thus, in this section we have outlined two kinds of localised distribution estimate — the localised kernel density estimate, and the LMPD. Use of the latter gives localised moment estimates in agreement with the intuitive definitions for all moments. Of course, one can also use localised mass point estimates as a basis for estimating any localised statistic, provided the statistic is well defined for a discrete distribution. For the remainder of the paper we use this method, for two main reasons — firstly, the agreement between the definitions in Section 3 and those derived from LMPDs seems an encouraging benchmark; and secondly, as shown in Section 5 deriving localised versions of non-moment-based statistics from LMPDs tends to give computationally simple results.

The practice of approximating a continuous distribution with a discrete one may seem strange, however, it is not unprecedented. For example, the technique of bootstrapping (Efron, 1979, 1981, 1982) relies on approximating a distribution from a sample $\{x_1 \ldots x_n\}$ as a mass point distribution with $w_1 = w_2 = \ldots = w_n = n^{-1}$. Indeed, a LMPD can be thought of as a generalisation of this distribution, in the same sense that localised kernel density estimates are spatial generalisations of ordinary kernel density estimates.

5. Applying the method to quantile-based statistics

An alternative to moment-based summary statistics are those based on quantiles, such as the median and the inter-quartile range. The sample-based estimates of these statistics tend to be more resistant to outliers than those for moment based statistics, and they play a key rôle in the exploratory data analysis methodology set out by Tukey (Tukey, 1977). In this section we consider the derivation of localised quantile-based summary statistics using LMPDs. A number of such statistics are listed in Table 1. Perhaps the least familiar of these is the quantile imbalance. This is based on the position of the median relative to the first and third quartiles, and measures the symmetry of the middle part of the distribution. It ranges from -1 (when the median is very close to the first quartile) to 1 (when the median is very close to the third quartile), and is zero if the median bisects the first and third quartiles. Essentially, this measures the degree of imbalance in the location of the median with respect to the quartiles considered by Brimicombe when proposing the normalised boxplot as an exploratory data analysis tool (Brimicombe, 1999). Unlike the skewness, it is not affected by the shape of the tails of the distribution, measuring only the shape of the distribution between the outer quartiles.

Note that all of these statistics are functions of the general p-quantile for various values of p. Thus, deriving localised p-quantiles from LMPDs is the key to obtaining the remaining statistics. Since we are working with LMPDs, we need the discrete distribution expression for the p-quantile, which is the minimum solution for q of the equation:

$$\Pr(x < q) = p. \qquad (19)$$

For a LMPD, we can write this expression as:

$$\sum_{x_i < q} w_i = p. \qquad (20)$$

This expression may be best understood if we label the x_i's in ascending order — and of course label the corresponding w_i's accordingly. Regard the left hand side of Eq. (19) as a function of q. This takes value of the sum of the set $\{w_1, w_2, \ldots w_J\}$, where J is the index of the largest x_i not exceeding q. This function jumps by an amount w_i each time q exceeds a value x_i. Thus, the left hand side only takes one of the n values w_1, $w_1 + w_2, \ldots w_1 + w_2 + \ldots + w_n$. (Note that the last of these is equal to one.) Unless one of these values happens to equal p, Eq. (20) has no solution. For some J we will have $w_1 + \ldots + w_J < p$ and $w_1 + \ldots + w_{J+1} > p$. A problem therefore arises: when this happens, it appears that the LMPD has no p-quantile.

We overcome this difficulty by extending the defining a p-quantile in this situation by interpolation. When $q = x_J$ we have $\Pr(x \leq q) = w_1 + \ldots + w_J = w_J^*$, say, and when $q = x_{J+1}$ we have $\Pr(x \leq q) = w_1 + \ldots + w_{J+1} = w_{J+1}^*$ If we were to assume that $\Pr(x \leq q)$ were a linear function between $q = x_J$ and $q = x_{J+1}$, rather than the discontinuous jump it actually is, then the solution to Eq. (20) would be:

$$q = x_J + (x_{J+1} - x_J)\frac{p - w_J^*}{w_{J+1}^* - w_J^*} \qquad (21)$$

This is just the standard linear interpolation formula. Finally, noting that $w_{J+1}^* - w_J^* = w_{J+1}$ the result may be simplified to:

$$q = x_J + w_{J+1}^{-1}(x_{J+1} - x_J)(p - w_J^*) \qquad (22)$$

The reader may wish to check that if $p = 0.5$, and all w_i's are equal to n^{-1}, then we obtain the standard expression for the sample median. Once again, we obtain an expression which can be thought of as the geographically weighted generalisation of as a global summary statistic.

Hence, via Eq. (22) we obtain localised versions of the median, the interquartile range and the quantile imbalance. We have thus used LMPDs to define a set of quantile based descriptive statistics to measure location (i.e. typical values of x), spread and symmetry of a variable of interest, x. These complement the moment-based descriptors derived in Section 4. Often, both types of statistic are of use: typically, the quantile based estimates are more robust to outlying values (as stated earlier), but the moment-based estimates tend to be smoother. We therefore suggest a set of six summary statistics that may be used to investigate spatial variates in univariate distributions, as set out in Table 2.

Note that this list is not exhaustive. For example, one could go on to consider kurtosis (based on the fourth moment), or local modes. However, we feel that this set of statistics provides a useful grounding for an exploratory analysis of local distribution shape.

Table 2
A typology of local summary statistics

	Location	Spread	Symmetry
Moment based	Mean	Standard deviation	Skewness
Quantile based	Median	Interquartile range	Quantile imbalance

6. Computational issues and visualization

In Sections 2 and 5 a methodology for deriving localised summary statistics is set out. However, for this to be of practical use consideration must also be given to issues of computation and visualization of these statistics. Clearly the above results would be of little value to applied geographers if they required impractical amounts of time to compute, or if they could not be illustrated effectively. In this section we outline the approaches we have adopted in both of these areas.

6.1. Computation

Computing locally weighted statistics via LMPDs will inevitably require a large amount of computation. As stated above, localised statistics can be regarded as mathematical functions defined over continuous regions on a map. However, in practice they will be evaluated over a regular lattice of points, usually a regular grid. Since the w_i's vary spatially, these must be re-computed for each point in the grid. As there needs to be a reasonably fine lattice to obtain reasonable results (typically between about 600 and 2000 points), this is generally the most time-consuming part of computation. The next most time consuming task is perhaps the sorting of the x_i's which is necessary to apply the localised p-median computations as outlined in Section 5. However, this task only has to be performed once for the entire lattice.

To compute the statistics in Table 2, it is worth noting that there is a large degree of cross referencing in their definitions — for example, the standard deviation definition refers to the mean, and the skewness refers to both the mean and standard deviation. Assuming the sorting of the x's and w_i's has already been carried out, all of these dependencies are shown in Fig. 8. Note that although the quantile imbalance depends on both the median and interquartile range, no arrows connecting these boxes are shown as the routine to compute this quantity needs only call on the p-quantile routine.

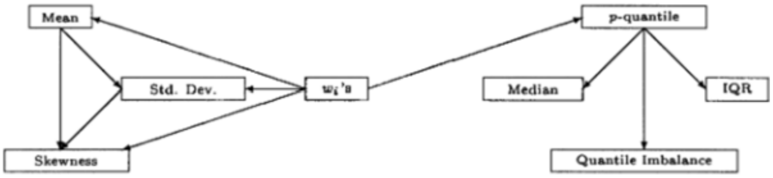


Fig. 8. Dependencies between local summary statistics.

In the diagram, if an arrow points from box A to box B, then the quantity in box A must be computed in order to compute the quantity in box B. As can be seen, everything depends on the w_i's, either directly or indirectly. The most effective order in which to compute all of the statistics is to ensure that for any statistic, all of the other statistics pointing to it have already been computed and stored as intermediate results. Note that this implies that it is better to compute all six statistics in Table 2 in a single procedure, rather than creating six stand-alone procedures, each requiring the w_i's to be computed independently. In fact, if the latter approach were taken, computing all six statistics would require more than six repeat calculations of the w_i's, as some of the procedures for more advanced statistics would call the more basic procedures. Here, we have carried out the coding of the localised statistic computation using the Lisp-Stat package (Tierney, 1990).

6.2. Visualization

Having computed localised statistics over a geographical study area, some consideration should be given to the method of visualizing the results. Since localised statistics can be represented as surfaces, or as abstract terrains, the obvious choices for visualisation are those of contour (isoline) maps, three-dimensional surfaces or unclassed choropleth grid maps (Tobler, 1973). All of these methods have a number of distinct advantages and disadvantages (Kraak & Ormeling, 1996). Although it is recommended that all of these techniques may be experimented with, here we have opted to use isoline maps. The main reason for this is that it is intended to compare several localised statistics — namely the six statistics suggested above — by applying the principle of small multiples (Tufte, 1990). This is interpreted here as showing small maps arranged in two rows and three columns — where splitting by row divides the display into moment-based vs. order based statistics, and splitting by column divides the display into measures of level, spread and symmetry. It has been suggested (Kraak & Ormeling, 1996) that isoline maps are the most suitable choice of visualization method for making such multiple comparisons.

7. An example

Here we apply the six localised statistics outlined in Table 2 to the house price data introduced in Section 1. The results are illustrated in Fig. 9. Here, all six indicators are shown as described above, visualized in the form of isopleth maps. Note that to reduce complexity in the multiple images, the place names and scale appearing in Figs. 3 and 4 are no longer shown. Note also that the mean and median share the same contour levels, as do the standard deviation and the interquartile range. This allows direct comparisons to be made.

Recall that the data comprise 1067 observations of house prices over a 1-year period for the county of Tyne and Wear in northeast England (Fig. 1). The local statistics can be used in a number of ways to explore these data. A choice needs to be made on where to provide an estimate of the local statistics. We may provide

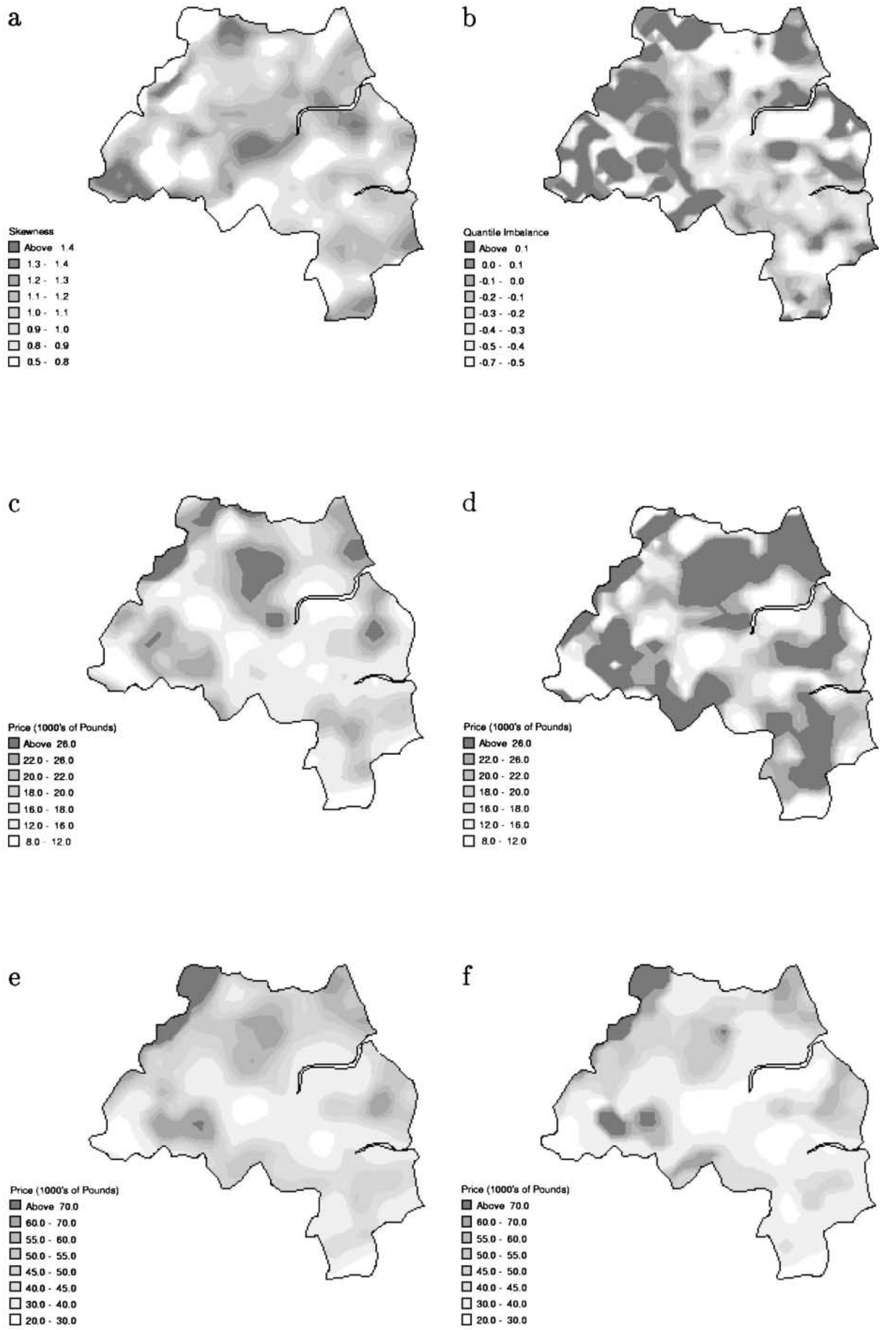


Fig. 9. Multiple maps of localised statistics. Maps are alphabetically labelled: (a) skewness; (b) quantile imbalance; (c) standard deviation; (d) inter-quartile range; (e) mean; and (f) median.

estimates at the points at which the data are sampled or at the mesh points of a regular grid. Alternatetively, we may provide estimates at some entirely different set of locations (for example, the centroids of postcodes or zip codes).

If it is desired to examine the price of a property in relation to others in the neighbourhood, the difference between its actual price and the local mean or median will give some indication of whether the house is locally more or less expensive. The local variance or interquartile range will give some indication of housing mix. The local skewness or quartile imbalance may give some insights as to whether the property is unusually expensive or not.

We may be interested in broader patterns. If this is the case, then we may wish to provide estimates of these local statistics at the mesh points of a regular grid — such data can then be conveniently read into some GIS software to provide a picture of the patterns across the region. This method has been used to generate some of the plots in this paper, and it obviates the need to use any interpolation routines to create a field from a set of irregularly spaced observations. Additionally, the user can then query some of the individual data for insights into why local variations in whatever is being summarised are taking place.

We may wish to provide estimates for some other set of locations. For instance, if we are interested in obtaining some measure of affluence for, say unit postcodes (the UK equivalent of the US zip code area), then we might evaluate the local mean house price at the centroid of each postcode. The UK Census of Population does not collect information on household income, and consequently there is a large geodemographic industry that attempts to estimate such data. Local estimates of housing costs could provide a useful proxy, and looking at the localised distributions of housing cost provides an indication of the degree of 'social mixing' in an area. Similarly, in epidemiological studies, it is sometimes useful to try to calculate the probability of a birth being stillborn relative to local estimates of deprivation (Dummer, Dickinson, Pearce, Charlton, Smith, & Salotti, 1998). Estimates of local housing cost may help to provide further information on local levels of deprivation in such studies.

Several experiments with different bandwidths were carried out on the house price data. The physical extent of the county is about 32 by 22 km. A bandwidth of 3 km gives a rather general picture of local house price trends, indicating a roughly north-west to south-east trend in mean house price — see Fig. 3. As it was desired to reveal variations in price between the settlements in the county, a bandwidth of 1 km appeared to provide an appropriate level of detail. However, there is no reason why several bandwidths may not be tested in order to reveal local variations at different spatial scales.

The map of the local means in Fig. 9e reveals some interesting patterns (see Fig. 1 for locations). The highest levels are to be found west of Seaton Burn. Recall that the local mean house price is around £44k. Examination of the data reveals two properties sold for £115k and £63k in the locality — the wide local standard deviation is evident from Fig. 9c. There is another high price locality around Forest Hall. One of the properties in this suburb was sold for £100k. Other areas of expensive housing are to the north end of Whitley Bay, Cleadon, Whickham and Rowlands

Gill. At the other end of the price scale we find areas of low cost housing around Chopwell where properties can be bought for as little as £20k. The west end of Gateshead is another area of low cost housing as is the south end of Hetton-le-Hole in the extreme south of the county. One regional pattern which is apparent among the local detail is that housing in the older industrial areas along the Tyne (Wallsend, Hebburn, Felling) and along the River Wear in Sunderland is generally of lower cost than housing in the more affluent commuter suburbs. The disparity is evident in comparing local means in Tynemouth (£56k) and Wallsend (£32k), two towns which are less than 7 km apart.

Examination of the local standard deviation map in Fig. 9c contrasts areas where the housing is consistent in price against other areas where there is wider variation. The older former industrial areas along the river have not only low cost housing but housing which is similarly priced. Housing is often cited as an example of a commodity where the distribution of prices is usually positively skewed. Skewness measures the direction in which outliers in a distribution tend to occur. In some areas of high priced housing (Whickham, Rowlands Gill, Cleadon) the local price distribution exhibits a negative skewness suggesting that in general local housing costs are high but that there are a handful of relatively inexpensive houses.

Revisiting the maps of local mean (Fig. 9e) and local median (Fig. 9f) we observe quite noticeable changes in local gradient. This is perhaps most noticeable in the high price area around Whickham. Local medians would appear to be more sensitive to local differences than are local means. With a 1 km bandwidth, the maps of local interquartile range (Fig. 9d) and quantile imbalance (Fig. 9b) pick up rather too much local noise and some further smoothing with a larger bandwidth is probably desirable.

8. Conclusion

There are a number of ways in which the ideas in this paper can be extended. For example, although the illustrations given here all apply to univariate distributions, the same principles could be applied in the multivariate case. Here, instead of considering distributions of mass points in one dimensional space, one could consider the m-dimensional case. This leads to the localised variance-covariance matrices, and consequently to the notion of localised principal components.

Another development could be that of nonparametric modeling. In this paper, emphasis has been placed on the exploratory aspects of localised descriptive statistics. However, it may be possible to consider localised estimates as a means of calibrating nonparametric models. For example, a model of the form:

$$z_i \sim N(f(u_i, v_i), g(u_i, v_i)) \qquad (23)$$

where (u_i, v_i) is a point in space, and z_i is some variable of interest, could be calibrated using localised estimates of mean and variance to estimate the unknown functions f and g. However, such estimates would be of limited use unless one was aware of the bias and standard error. In situations like this, the use of bootstrapping methods (Efron, 1979, 1981, 1982) is likely to be of help. It is intended that this will be the subject of further research.

The technique also raises some questions regarding the choice of bandwidth. Since the emphasis in this paper has been exploratory, choice of bandwidth has been rather arbitrary. The choice of 1 km in the examples seemed a realistic distance for considering local variations in house prices. However, in some situations a bandwidth may not readily suggest itself, and 'automatic' approaches to bandwidth choice may be helpful. This itself raises another question — should one expect all localised descriptive statistics to have the same optimal bandwidth? Again, this will be the subject of future research.

In conclusion, a generalised approach to localised statistics has been proposed. Through the idea of localised distributions, and in particular the localised mass point distribution, a large number of localised statistics may be created. As well as specific examples, this paper provides a framework for creating arbitrary localised statistics. With geographically weighted regression (Brunsdon et al., 1996) an approach for handling geographical nonstationarity in regression models was proposed. It is hoped that the framework set out in this article, together with the extensions suggested above, will lead to a very general approach to handling geographical nonstationarity in many aspects of statistical modeling and data exploration.

References

Brimicombe, A. J. (1999). Small may be beautiful, but is simple sufficient. Geographical and Environmental Modelling, 3, 9–33.

Brunsdon, C. (1995). Estimating probability surfaces for geographical points data: an adaptive kernel algorithm. Computers and Geosciences, 21, 877–894.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical Analysis, 28, 281–289.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (1998). Spatial nonstationarity and autoregressive models. Environment and Planning A, 30, 957–973.

Burrough, P. (1986). Principles of geographic information systems for land resources assessment. Oxford: Oxford University Press.

Dummer, T., Dickinson, H., Pearce, M., Charlton, M., Smith, J., & Salotti, S. (1998). Stillbirth rates around the nuclear installation at Sellafield, North West England: 1950–1989. International Journal of Epidemiology, 27(1), 74–82.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1–26.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika, 68, 589–599.

Efron, B. (1982). The jacknife, the bootstrap and other resampling plans. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Ehrenberg, A. (1982). A primer in data reduction. Chichester: Wiley.

Fotheringham, A., Charlton, M., & Brunsdon, C. (1997). Two techniques for exploring non-stationarity in geographical data. Geographical Systems, 4, 59–82.

Fotheringham, A., Charlton, M., & Brunsdon, C. (1998). Geographically weighted regression: a natural extension of the expansion method for spatial data analysis. Environment and Planning A, 30, 1905–1928.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford University Press, Oxford, UK.

Kraak, M.-J., & Ormeling, F. (1996). Cartography — visualization of spatial data. Essex: Longman.

Krige, D. (1966). Two-dimensional moving average surfaces for ore evaluation. Journal of South African Institute of Mining and Metallurgy, 66, 13–38.

Lam, N. (1983). Spatial interpolation methods: a review. The American Cartographer, 10, 129–149.

Matheron, G. (1973). The intrinsic random functions and their applications. Advances in Applied Probability, 5, 439–468.

Openshaw, S. (1984). CATMOG 38: the modifiable areal unit problem. Norwich: Geo-Abstracts.

Openshaw, S. (1991). Developing spatial analysis methods for GIS. In D. Maguire, M. Goodchild, & D. Rhind, Geographical information systems: principles and applications (pp. 389–402). London: Longman.

Ripley, B. (1981). Spatial statistics. New York: Wiley.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. London: Chapman and Hall.

Tierney, L. (1990). LISP-STAT: an object oriented environment for statistical computing and dynamic graphics. Chichester: Wiley.

Tobler, W. (1973). Choropleth maps without class intervals? Geographical analysis, 3, 262–265.

Tobler, W. R. (1989). Frame independent spatial analysis. In M. F. Goodchild, & S. Gopal, The accuracy of spatial databases (pp. 115–122). London: Taylor and Francis.

Tomlin, C. (1990). Geographic information systems and cartographic modelling. New Jersey: Prenctice Hall.

Tufte, E. R. (1990). Envisioning information. Cheshire, Connecticut: Graphics Press.

Tukey, J. (1977). Exploratory data analysis. New York: Addison-Wesley.

Wand, M., & Jones, C. (1995). Kernel smoothing. London: Chapman and Hall.