Smoothing of Standardised Mortality Ratios

A Preliminary Investigation

National Statistics Methodological Series No 35

Anca Carrington
Patrick Heady
Martin Ralphs
Nargis Rahman
Bruce Mitchell

Office for National Statistics
1 Drummond Gate
London SW1V 2QQ
Tel: 020 7533 9233
Fax: 020 7533 9292

July 2007

ISBN 978-1-85774-653-2.

Contact points

For enquiries about this publication contact:

Martin Ralphs on 020 7533 6383

Email: martin.ralphs@ons.gov.uk

For general enquiries, contact the National Statistics

Customer Contact Centre on 0845 601 3034

(minicom: 01633 812399)

Email: info@statistics.gov.uk

Fax: 01633 652747

Post: Room 1015, Government Buildings, Cardiff Road, Newport, NP10 8XG

About the Office for National Statistics

The Office for National Statistics (ONS) is the government agency responsible for compiling, analysing and disseminating economic, social and demographic statistics about the United Kingdom. It also administers the statutory registration of births, marriages and deaths in England and Wales. The Director of ONS is also the National Statistician and the Registrar General for England and Wales.

A National Statistics publication

National Statistics are produced to high professional standards set out in the National Statistics Code of Practice. They are produced free from any political influence.

Table of contents

List of tables

List of figures and maps

Acknowledgements

The authors would like to thank Allan Baker, Clare Griffiths and Barbara Toson of Social and Health Analysis and Reporting Division for their assistance and feedback throughout the duration of this project. We would also like to thank Professor Nicky Best of Imperial College London and Simon Compton and Marie Cruddas of the Methodology Directorate at the Office for National Statistics for their constructive criticism and feedback on this report.

Management Summary

The objective of smoothing methods is to enhance the visibility of underlying trends in noisy data. An underlying component - free of error – can be made more visible or approximated via smoothing. This report describes the results of a preliminary investigation into the use of spatial and temporal smoothing methods to remove noise from ward-level standardised mortality ratios that the ONS has published at ward level (http://www.statistics.gov.uk/statbase/Product.asp?vlnk=14359).

The investigation had two main objectives: first, to establish whether spatial data smoothing methods appeared to provide more stable estimates of underlying patterns of mortality than the approach currently undertaken by Social & Health Analysis & Reporting Division (SHARD); second, to identify which, if any, of the approaches was a suitable candidate for application in the medium to long term and should be taken forward.

The main focus of the report is the establishment of the theoretical basis for smoothing, derivation of performance criteria to assess smoothing stability and predictive power and the formal specification of basic methods for data smoothing across time, across space and across both time and space, using a range of configurations for the smoothing procedures. The methods thus developed are then implemented using a combination of statistical software packages and applied to ward-level mortality data supplied by SHARD. An assessment of performance is provided and the robustness of the selected candidate methods is evaluated. The method currently implemented by SHARD and the raw, annual unsmoothed data provide benchmark cases against which performance is compared.

The key finding of the analysis is that, for the data set under consideration, it is beneficial to apply some form of smoothing to annual SMR series if the objective is to reveal the underlying pattern in the data by reducing noise. The highest gain in stability was obtained by smoothing data in both time and space, with no loss of predictive power. Smoothing over time alone induced less stability than smoothing over time and space, but more than spatial smoothing alone. The method currently used in the ONS for computing five-year SMRs at ward level belongs to the family of smoothing solutions that perform best across time. Further gain is offered by space-time smoothing, but if this is applied to the SMR then it comes at the price of spatially varied performance. However, separate smoothing of the SMR death and population counts appears to counter this problem. We recommend that, for the present, the time smoothing solution that is currently deployed is sufficient. However, there is likely to be considerable potential in applying a hybrid space-time method in future, particularly if we can build on the basic approaches implemented here. It would also be worthwhile to investigate other smoothing methods (for example Bayesian smoothers) and compare their performance against the simpler approaches deployed here using a cross-validation framework.

It is important to note that this recommendation is based on the evaluation of the performance of similar smoothing solutions over three years and that, if more five-year SMR measures become available, their direct evaluation will be required.

The findings we report are based on the exploration of a limited number of options, and our conclusions are therefore preliminary. Additionally, the evaluation has concentrated on global performance measures, but some factors (such as the choice of boundary

geographies) may have a local impact on performance and this should be evaluated. We recommend that further work should be undertaken to look in more detail at the analytical impacts and practical implementation issues arising from the use of space-time smoothing techniques with a view to deploying such more sophisticated methods to produce experimental statistics for publication in the future.

1: Introduction and Scope

This is a report on the findings of a preliminary study on methods of data smoothing over space and time and their application to ward-level standardised mortality ratios. The study was undertaken by the Spatial Analysis Centre in ONS Methodology Directorate and was commissioned by Social and Health Analysis and Reporting Division (SHARD).

The report is divided into five chapters. In this first chapter we set out our objectives and specify the problem. In chapter two we review the literature on smoothing. Chapter three sets out the key theoretical concepts on which smoothing is based, specifies a framework for evaluation and defines the methods of smoothing data over time and space that we have constructed and applied. In chapter four we present our performance evaluation for the smoothing solutions investigated. Chapter five sets out our conclusions and recommendations for future work on this topic.

The content of this report is limited by the exploratory nature of the work. Our focus is on:

1. The definition of underlying concepts and of criteria for the evaluation of smoothing methods, and their development and formal specification in algebraic terms.
2. The formal specification of some basic candidate methods for smoothing area-based data over time and space.
3. The presentation and evaluation of a set of initial results obtained by implementing these basic methods and applying them to standardised mortality ratios at ward level using area level data supplied by SHARD.

The investigation had two main objectives. The first one was to establish whether spatial data smoothing methods (and hybrid space-time smoothers) appeared to provide more stable estimates of underlying patterns of mortality than the time-smoothing approach undertaken already by SHARD. The second was to identify which, if any, of the approaches was a suitable candidate for application and should be taken forward.

This work represents a first step towards the development of effective methods for reducing the instability present in the ward-level standardised mortality ratios published by the Office for National Statistics. The report does NOT constitute a manual defining best practice for the application of smoothing methods in the context of standardised mortality ratios. A detailed treatment of the practical considerations of applying these methods and best practice in implementation is beyond the scope of this initial study.

Specification of the problem: instability of mortality measures for small areas

Mortality can be measured in a number of ways (Smith, 1992): by a crude rate, calculated as (deaths/population) * 1000; using direct standardised rates (i.e. specific rates from target population applied to standard population) or a standardised mortality ratio (henceforth abbreviated to SMR).

One key indicator used in the reporting and analysis of geographical variation in mortality is the independent, area-specific SMR for various territorial units (counties, census tracts, electoral wards), for which count data on cases and population at risk are available, making maps a convenient and accessible summary of numerical tables (Elliott, Martuzzi and Shaddick, 1995; Best et al., 1999).

This approach is less than satisfactory when dealing with small areas, where estimates based on few deaths and a small number of people at risk are inherently unstable (Waller, Carlin and Xia, 1997) and create the impression of spurious geographic variation (Veugelers and Hornibrook, 2002).

To avoid potential problems from conclusions being drawn on the basis of unstable estimates based on a limited set of observations, one of the following solutions could be chosen:

- Mapping the significance level of estimated risk rates for each area rather than the risk figure itself. In this case, no estimation of the size of the risk itself is given, but whether or not it is statistically significant is conveyed;
- Shading maps or using a combination of colour and symbology to display both the size of the estimated risk (e.g., SMR) and its statistical significance (for example its corresponding rank in the set of p–values). This requires very careful selection of colours and symbols and is difficult to achieve effectively in practice;
- Smoothing the raw values and therefore reducing the error of estimation over the entire map (Elliott, Martuzzi and Shaddick, 1995);
- Mapping exceedance probabilities over critical threshold values calculated using Bayesian methods.

The smoothing solution brings benefits that go beyond enhancing the clarity of the information contained in a map of SMRs. Firstly, smoothing helps stabilise rates based on small numbers at the desired level of spatial disaggregation; secondly, it reduces noise caused by different population sizes used in the calculation of rates. The ultimate gain is the increased ability of the user to discern more clearly systematic patterns in the spatial variation of underlying risk (Waller and Gotway, 2004; Kafadar, 1996).

While the estimate of risk in any single area is optimal when location is not seen as relevant and independence across space is assumed, it is possible to derive improved estimates of the relative risk by building estimators that take account of spatial dependence and borrow strength in space. This means a change of approach from regarding each area as a self-contained entity to taking into account the ensemble of areas under study and the high probability that underlying patterns of mortality may well cross small area boundaries. This refinement is important when using SMRs as measures of the underlying area-specific risks, since in their raw form they do not allow for spatial correlation 'which may be induced through dependence on shared but unmeasured risk factors which themselves vary systematically in space' (Best et al., 1999: 132). Spatially smoothed estimates are, therefore, more appropriate for the judgement of geographic variation than those which take no account of spatial dependence (Veugelers and Hornibrook, 2002).

It is important to recognise that there are also disadvantages associated with smoothing. Firstly, smoothing introduces autocorrelation (i.e. correlation among neighbouring values) in the data, and may be seen as replacing unstable estimates with correlated ones. Secondly, the differences between the smooth and the raw values may be a source of concern for users. As Waller and Gotway (2004) warn, '[M]any people are leery of statistically adjusted numbers, particularly if money or power is to be allocated based on them' (p.97). The publication of both raw and smoothed values is one way to overcome this problem but requires sufficient guidance for users to ensure appropriate use of the statistics. In fact, all of the statistical indicators of mortality for small areas published currently by the ONS incorporate some level of smoothing. On balance, given how misleading inferences based on maps from raw rates can be, smoothing is regarded as preferable if the extraction of underlying pattern is the desired outcome (Waller and Gotway, 2004).

Objectives and project framework

Objectives

In this project, we focus on the smoothing of SMRs, across time, across space and across both time and space. Our objectives are to:

- Define a selection of simple time and space smoothing techniques and apply them to ward-level SMRs;
- Evaluate the performance of the simple smoothing solutions defined;
- Identify the most suitable simple smoothing solution for the SMR data under analysis;
- Indicate the direction for future work in relation to smoothing SMRs.

We chose to concentrate on simple smoothing techniques for two reasons. Firstly, as this is a first project devoted to the exploration of smoothing solutions in the Office, we aimed to start by clarifying the basic ideas of smoothing and on conveying the essence of the technique, best reflected in the most simple of the smoothers. Secondly, we envisage that the smooth values will be used in an exploratory way (in this case to provide visual clues as to the underlying geographical distribution of mortality in England and Wales), for which purpose 'a smoother that can be calculated quickly without complex estimation rules, difficult or tedious programming, and specification of extra parameters is preferred' (Waller and Gotway, 2004: 98). More complicated solutions, closely associated with inferential statistics (such as Bayesian ones) are beyond the scope of the current project.

Finally, we note that while this report deals exclusively with the smoothing of SMRs, the techniques applied and evaluated here may also be suitable for smoothing other types of data.

Project Framework

The framework of this project is determined by the nature of the data analysed (SMRs) and of the techniques applied to them (smoothing). We discuss each in turn briefly in this introduction, but the reader is referred to Appendix A for a more detailed explanation of SMRs and their calculation and to part two of this report for a full discussion of smoothing.

By definition, the Standardised Mortality Ratio (SMR) is the ratio of the observed to the expected number of deaths in an area, multiplied by 100:

$$SMR = \frac{\text{observed deaths}}{\text{expected deaths}} \times 100 \quad (1)$$

Observed deaths are those which actually occurred in the local population in the year in question. Expected deaths are those which would be expected given the age structure and size of the local population, if it had the same age-specific death rates as the reference population.

In this project, the local population is that of each ward, while the reference population is that of England and Wales (male and female, below the age of 85). Observed deaths are calculated for each ward as the total deaths (i.e. the sum of deaths, across age groups up to 85, and gender).

The smoothing techniques applied here are linear in nature and, in essence, involve simple and weighted averaging over predefined neighbourhoods of values. The methods we apply belong to a spectrum of techniques detailed in the "Data and Techniques" part of chapter 2. The term "neighbourhood" in this context requires some clarification. In the context of space smoothing, such neighbourhoods follow the conventional idea of contiguous areas in space. However, in the time-smoothing sense a "neighbourhood" can also include past values of the target area in time. For space-time smoothing, both types of neighbours might be included.

Technical Implementation

No single software package available in the Office for National Statistics offered all of the tools necessary for the implementation of the smoothing techniques discussed here. The main constraint comes from the need to capture and use geographical information in an analytical context. As a consequence, more than one package has been evaluated and a combination of packages has been used to achieve the results reported. Specifically, the Geographical Information System (GIS) package ArcGIS 8.3 was used to prepare data and produce maps and the spatial analysis tool GeoDa to compute the information on location and spatial relationships. This information was then transferred to MATLAB 7.1, whose strength in dealing with very large sparse matrices made it the ideal environment for the final computations[1].

[1] *S-Plus 2000* (including *S+ Spatial Stats*) handled the computations equally well, but proved to be around 100 times slower on equivalent size data sets than *MATLAB*.

2: Smoothing – A Literature Review

Smoothing is the application of specific procedures in order to remove noise from data. This can be achieved over time, across space or a combination of the two.

The emphasis in this review is on the types of smoothing techniques available, with links to the relevant theoretical underpinnings and practical implementation solutions. The aim is to offer an introduction to smoothing - why it matters, what it can offer and what the main ways of approaching it are, as presented in the academic literature. The review offers an outline answer to all these questions and therefore does not contain a detailed presentation of the theoretical background.

The remainder of this chapter is structured in four parts. The first discusses the motivation behind temporal and spatial smoothing. This is followed by an introduction to the techniques available for different types of data. We move on to discuss the implementation of these and finally we draw some conclusions.

Motivation

Sharp variations in data, from one moment in time to the next or from one area to another may make it hard to detect meaningful underlying patterns. These variations are, partly, real changes in the underlying phenomenon measured and, partly, noise. In other words, we can view the data as being made up of two component parts:

Data = Smooth + Rough (2)

The 'rough' part - or the noise - can be due to measurement or sampling error or to underlying random variation. By removing the noise, smoothing works to uncover (the term "retrieve" is commonly used in the literature) the true underlying component (Tukey, 1977: 235).

Formally, a random process measured in a number of locations *s*, across *d* dimensions, over a region of analysis *D* $\left\{Z(s): s \in D \subset \mathbb{R}^d\right\}$ can be decomposed conceptually into a smooth or noiseless component and a white noise, randomly varying error component:

$$Z(s) = S(s) + \varepsilon(s), \quad s \in D \qquad (3)$$

Smoothing is about retrieving the noiseless part of the process under analysis, which is represented in equation (3) by $S(s)$.

Smoothing over time has been well established in the exploratory analysis of time series since Tukey (1977) as a means of retrieving large-scale or long-run behaviour (Mills, 1990). The benefits from spatial smoothing are equally well documented in the literature: the production of smoothed maps from which spatial patterns of mortality risks can be monitored over time (MacNab and Dean, 2001); reduced variability of small area estimates (Cowling et al., 1996); insight into geographic patterns that may be hidden in noisy data (Kafadar, 1996) and smoothed estimates that allow for better judgement of the geographic distribution (Veugelers and Hornibrook, 2002: 102). Spatial smoothing is viewed as suitable when data are subject to a high degree of variation from many error sources such as counting and recording errors (Kafadar, 1996) or when 'data correspond to regions and apply to people' (Kafadar, 1994: 420). The latter refers to the common practice of reporting statistics about individuals in aggregate form at area level and certainly applies in the case of SMRs.

According to Kafadar and Horn (2001), smoothing is a flexible exploratory technique that can be used in order to:

- magnify an underlying trend;
- detract attention from outliers[2];
- examine patterns in residuals after the elimination of the trend;
- minimise the effect of aggregation of data;
- reveal relationships between response and explanatory variables (p.1).

This works for a wide range of data types, from long aggregated time series to (spatially disaggregated) regional data at a moment in time. As far as time series are concerned, their exploratory analysis is often driven by the desire to focus on large-scale or long-run behaviour. The presence of noise can make this task a very difficult one, in particular when these series are very short ones, which is often the case with sub-national data. Furthermore, moving from data across time to data over space, as Kafadar (1996) argues, '[G]eographically-defined data are often amenable to smoothing, since data in one region are likely affected, to a greater or lesser extent, by data in neighbouring regions. Smoothing borrows neighbouring information in a flexible way to permit exploratory analyses and provide indications of possible patterns that one might otherwise find difficult to detect.' (p.2557).

Data that have been smoothed can be used to produce maps from which geographical patterns previously hidden by noise can be uncovered and monitored and to obtain small area estimates with reduced variability (Cowling et al., 1996).

Such techniques have been widely applied to small area health-related data such as disease rates (Kafadar, 1996; Best et al., 1999) and mortality risks (Kafadar, 1999; MacNab and Dean, 2001), as well as to environmental data (e.g. Grillenzoni, 2003).

Data and Techniques

Smoothing techniques can be applied to either point or area data, at one moment in time or along a time series. While the techniques for analysing each type of data have evolved separately, they remain largely interchangeable, with careful adaptation to the data available.

Data subject to smoothing

Whether at a moment in time or part of a time series, point data refer to measures of attributes for individual observations at identified locations. Area data are the outcome of aggregating individual data to a set of boundaries (as a sum, count, mean etc.).

[2] However, according to (Kafadar, 1999), a good smoother ensures that unusual values stand out clearly in the residuals, not in the smooth.

Smoothing techniques

Different smoothing techniques produce different results when applied to the same data. Furthermore, within a given technique, the outcome of smoothing is closely linked to the choice of smoothing parameters (time span, area covered, specific weights): the larger the degree of smoothing, the smaller the variance of the smoothed value and the larger the bias (the amount that the smooth value is different from the raw value). The reduction in variance at the cost of increased bias in the smoothed value is the fundamental trade-off between bias and variance in all smoothing methods.

Smoothing techniques are either linear or non-linear. ***Nonlinear smoothing*** relies on combinations of medians of nearby data values (techniques include head-banging, re-smoothed medians and median polishing). The non-linear approach responds more quickly to abrupt changes in value and is not influenced by outliers that are not supported by neighbouring values. However, given that non-linear functions have linear local approximations and that well-chosen linear smoothers can therefore provide a good approximation of the more sophisticated non-linear forms, we concentrate on linear smoothers in the remainder of this review.

Linear smoothing techniques work using means, have a less sharp reaction to abrupt changes in the data and cover a wide range of smoothing solutions on which we concentrate below. For ease of presentation, these methods can be grouped into four main categories, which we discuss in turn:

1. Simple linear smoothing (running means)
2. Kernel smoothing
3. Regression smoothing
4. Bayesian smoothing

Note that these families of methods are not mutually exclusive, as kernels can be used in regression, regression can be used for Bayesian smoothing etc.

Simple linear smoothing (running means)

Linear smoothing applied to a target variable (y_i) produces a smooth value ($\tilde{y}_i$) that is a weighted average of a set of relevant values (y_j):

$$\tilde{y}_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}}, \qquad (4)$$

Decisions have to be made about what observations (j) to include in the smooth and with what weights.

Linear smoothing can be applied over time (in which case y_j are the past values for unit i), over space (in which case y_j are the observed current values for a set of neighbours of unit i) and over both time and space (in which case y_j are the observed current and values for i and a set of its neighbours). The smooth estimate for area i can include the raw value for that area, but this is not a requirement for the smooth to be computable (when the raw value is missing, the smooth value can be interpolated using information about the past and / or neighbours).

As far as smoothing across space is concerned, areas *j* belong to a predetermined neighbourhood (e.g. areas that share common border with *i*; areas within a certain distance of *i* etc.). The simplest specification of the weights that link the target area and its neighbours is the binary contiguity one, specified as:

$$w_{ij} = \begin{cases} 1 & neighbours \\ 0 & otherwise \end{cases} . \qquad (5)$$

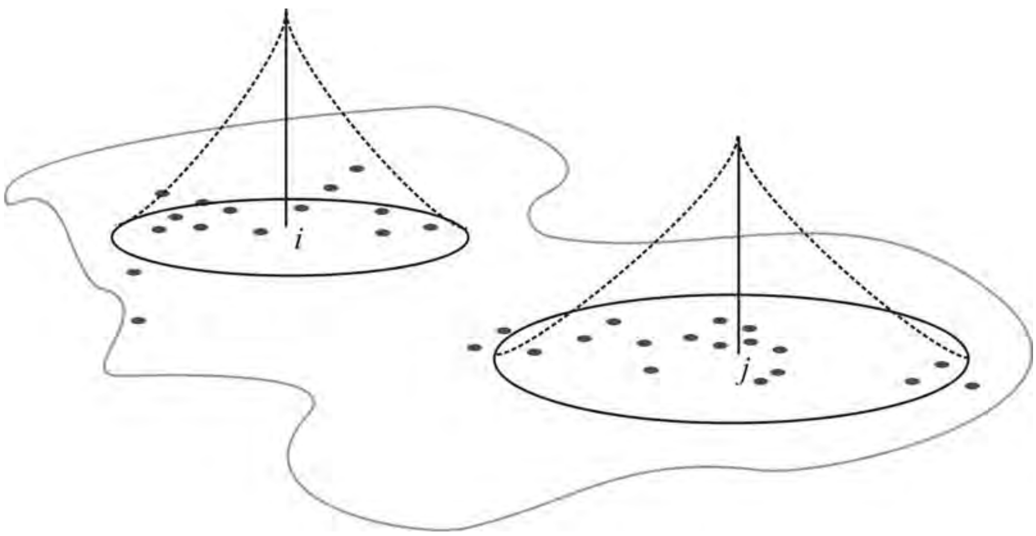
More elaborate spatial weights can be specified in ways that allow for differences in the extent of influence between different pairs of neighbours (e.g. as a function of distance between centroids of areas, length of shared boundary etc., Getis and Aldstadt, 2004, Griffith, 1996).

The choice of relevant time horizon, neighbours and weights is arbitrary. The larger is the proportion of values included in the smooth (i.e. the number of non-zero weights), the smoother and more stable is the resulting surface, but also the greater is the bias.

Kernel Smoothing

The basic idea of kernels is that for each point of interest there a 'bump of influence' (Fotheringham, Brunsdon and Charlton, 2002:59) that relates it to neighbouring values. In spatial smoothing this bump takes the form of a three-dimensional shape like a pointed hat or a tee-pee tent. The user specifies the parameters which define the shape and size of this mathematical 'tent'. Figure 2.1 illustrates how this works in practice.

Figure 2.1 – Simple circular kernel functions placed around two data points i and j in space. The diameter of the 'tent' determines how many data points are included in the smoothing operation, while the gradient of the 'tent' defines how the influence of more distant points is taken into account by the kernel estimator.



The shape and size of the 'tent' is what the analyst specifies when setting up the kernel smoothing function. Kernel smoothing is a form of non-parametric smoothing that works by placing 'tents' over individual observations and specifying the weights w_{ij} between observations i and j as a function of (geographical) distance, with more distant observations having less influence than near ones. The weighting function used determines the shape of the sides of the tent. In mathematical terms, the tent is actually a probability density function[3], typically unimodal and symmetrical and usually referred to as a 'kernel'). In this approach the weights take the following general form:

$$w_{ij} = \begin{cases} K(d_{ij}/h) & \textit{kernel neighbourhood} \\ 0 & \textit{otherwise} \end{cases} \tag{6}$$

where h is the smoothing parameter (often referred to as the 'bandwidth', the measure of the distance-decay in the weighting function), K is a kernel function and d_{ij} is the distance in space between i and j.

The simplest way of specifying the neighbourhood in this context is by using a kernel centred at the point of interest. For spatial smoothing, this window takes the shape of a disc. The points falling within that window (disc) are the ones whose values are included in the calculation of the smooth value. The weights allocated to these points can be equal or differentiated (for example according to their distance from the target point i, as in Figure 2.1).

[3] A three-dimensional one, in the case of spatial smoothing.

There are many different ways of defining the weighting function for kernel smoothing. For instance, the case of equal weights can be formally written as

$$w_{ij} = \begin{cases} 1 & \text{if } d_{ij} < d \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where *d* is the width of the window (or the radius of the disc), d_{ij} is the distance between point *i* and one of its neighbours (*j*) and w_{ij} is the weight given to the value at point *j* in calculating the smooth value at point *i*. This specification is simple but has the problem of discontinuity because the neighbourhood thus defined ends abruptly at the edge of the window.

One way of addressing this problem is by specifying the weights as a continuous function of the distance between any point and its neighbours, d_{ij}. For example, the weights could be specified as:

$$w_{ij} = \exp\left[-\frac{1}{2}(d_{ij}/h)^2\right]. \quad (8)$$

Here, the weight given to data points decreases according to a Normal curve as the distance between *i* and *j* increases. For data points far away from *i* the weights are thus very close to zero since [$\lim_{p\to\infty}\frac{1}{e^p} = 0$]. In practice this means that very distant points are effectively excluded from the calculation of the smooth value.

Weights might also be derived from a bi-square kernel, as in this example:

$$w_{ij} = \begin{cases} \left[1\text{-}(d_{ij}/h)^2\right]^2 & \text{if } d_{ij} < d \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Here, the weights vary with distance according to the function $\left[1\text{-}(d_{ij}/h)^2\right]^2$, but we have also specified that this should only apply when the distance between i and j is less than the fixed distance *d*. Where $d > d_{ij}$, points are therefore excluded from the calculation. This differs from equation (7) in that we have formally defined a weighting function to use when $d < d_{ij}$ rather than simply including all of the data points that lie within the cut-off distance.

Spatial kernels that have a small bandwidth have a steep weighting function (so that the 'tent' is tall and narrow) and produce a rougher surface than those with a large bandwidth. Once more, the outcome of smoothing depends on the choice of smoothing parameter. Making a sensible choice for the value of this parameter is thus of paramount importance. Several selection criteria are available in the literature.

The bandwidth can be chosen arbitrarily, using some rule of thumb or on the basis of cross-validation. For instance, Fotheringham, Brunsdon and Charlton (2000) propose a maximal smoothing rule for choosing *h* as a function of the sample standard deviation (*sd*) and the sample size (n):

$$h_{\max} \approx \frac{1.144sd}{\sqrt[5]{n}} \quad (10)$$

On the other hand, Simonoff (1996) recommends an exploration based on the use of a wide range of smoothing parameters, including locally varying ones (p.72).

Most robust but also most costly (in terms of time and computing power) is the cross-validation approach. This involves the calculation of a cross-validation score (CV) at various bandwidths and the selection of that bandwidth (*h*) that corresponds to the lowest CV score:

$$CV = \sum_{i=1}^{n} \left[y_i - \tilde{y}_{\neq i}(h) \right]^2 \qquad (11)$$

where $\tilde{y}_{\neq i}(h)$ is the smooth value at point *i* calculated without using the observed value at that point. The bandwidth thus chosen may not be the most suitable one for another data set. In contrast, an arbitrarily chosen bandwidth can be applied to any number of data sets, but it may never be optimal.

Furthermore, spatial kernels can be fixed or adaptive. When the smooth value is calculated on the basis of the values of those points that fall within the coverage of a fixed kernel (with a disc base), the number of points included will vary from one area to another, so the smooth values will be obtained on the basis of a varying number of points. Adaptive spatial kernels vary the bandwidth according to the variation in the density of the data points across space: a smaller kernel where the points are dense and a larger one where data are sparse similar to the nearest *k* neighbours idea. The kernel can allocate equal weights or ones that vary with distance from the target point.

The consensus in the literature is that ease of computation and properties of the smooth values are more important than the choice of kernel (Simonoff, 1996).

Regression Smoothing

Regression smoothing uses local linear and polynomial estimation methods to fit an optimised mathematical function to the input data. The relationship between a target variable *y* and one covariate *x* takes the following form:

$$y_i = f(x_i) + \varepsilon_i \qquad (12)$$

This specification – which can be generalised to include more than one covariate - can be approached parametrically (where the function *f* is specified in terms of parameters, e.g. $f(x_i) = a + bx_i$) or as a non-parametric regression (where the function *f* is left unspecified) and used to generate smooth estimates.

Smoothing can also be achieved in a semi-parametric approach. Across space, this involves estimating the smooth value of the target variable using information on both spatial coordinates (*u*,*v*) and a set of covariates (*x*), as in:

$$y_i = f(u_i, v_i) + \sum_k \beta_k x_{ik} + \varepsilon_i \qquad (13)$$

where no assumptions are made about the functional form of *f*. At point (*u*,*v*) smoothing is achieved by computing a weighted average of the residuals with the greatest weight assigned to observations closest to (*u*,*v*) (Müller, Stadtmüller and Tabnak, 1997, Fotheringham, Brunsdon and Charlton, 2000:179). This approach combines traditional regression analysis with kernels.

Regression smoothing can also be achieved using spline functions. These are polynomial functions that can have a very simple form locally, while remaining globally flexible. They are also able to accommodate heterogeneous relationships across time and space (Simonoff, 1996). Spline smoothers usually fit different polynomials in different segments (pieces) of the data, while imposing constraints to ensure that smoothness is also present in the move from one segment of the data to another. A function that fits to subsets like this is usually referred to as a 'piecewise' polynomial function. Segments can be pre-specified or found adaptively from the data.

A widely used geo-statistical technique known as 'kriging' is also part of the family of regression smoothers. Kriging produces spatial estimates as a weighted linear combination of the data, but in this case the weights are optimally derived via regression, on the basis of minimised error variance (Cressie, 1993).

Finally, wavelet smoothing is a spatially adaptive procedure based on the assumption that the smooth value is a linear combination of a limited number of basic functions, some of which have non-zero parameters (Kafadar and Horn, 2001:3).

Bayesian Smoothing

In empirical Bayes spatial smoothing (designed specifically for mapping disease rates), the weights depend upon the specification of a prior probability distribution of the variable of interest (Kafadar and Horn, 2001:3). Empirical or hierarchical Bayes smoothing models are based on the 'borrowing of strength' property. In this framework, the smoothed value is a weighted average of the original value and the global or local mean of the data; the former is favoured for a highly stable rate and the latter when the rate is based on few person years (Kafadar, 1994: 426). In other words, the amount of shrinkage towards the global or local mean is determined by the reliability in each particular territorial unit (region, area, ward). Known explanatory variables can be incorporated in this approach, by making the mean of the prior distribution a function of the covariates (Elliott, Martuzzi and Shaddick, 1995:152). Banerjee, Carlin and Gelfand (2003) provide a useful review of the hierarchical Bayesian framework for spatial data.

Relative Performance

Comparisons across smoothers suggest that 'relatively smooth underlying surfaces subjected to Gaussian noise benefit most when linear smoothing is applied, but unusual values, sharp features and non-Gaussian noise render them less efficient than non-linear smoothers' (Kafadar, 1999: 3173-74). Kafadar (1994) concludes, on the basis of numerous simulations, that weighted averages (with weights defined inversely proportional to squared distance) and the non-linear headbanging method are likely to be the most successful smoothers in practice. Meanwhile, Simonoff (1996) stresses the benefits of local polynomial estimators, as easy to interpret as a generalisation of linear regression, with a kernel-like nature that makes possible the use of refinements not available otherwise.

Unfortunately, as the above set of academic opinions suggests, there is not a universal solution to the question of which smoothing method to apply. The evaluations of alternative smoothing solutions reported in the literature are based on simulation results. In practice, appropriate frameworks for evaluation need to be designed and implemented given the application and data context and evaluating the performance of each option against real data. We present our specific approach to evaluation in the context of area-level SMR data in the third chapter of this report.

Implementation

The implementation of smoothing methods is complicated by the lack of an "off the shelf" technical solution that provides a single point of access for the implementation of smoothing. When spatial aspects are taken into account, the use of a combination of Geographical Information Systems (GIS) and statistical packages is necessary.

The most frequently used statistical package for this type of application seems to be S-Plus (more commonly its freely available version R) (loess, age-specific rates smoother, weighted head-banging, in Kafadar, 1999; spatially smoothed multilevel modelling of life expectancy, in Veugelers and Hornibrook, 2002; kernel and spline smoothing, Simonoff, 1996). Another useful package is Mathematica (MCMC, in Knorr-Held and Raβer, 2000). More powerful in its ability to handle large matrices (which are required in the quantification of spatial interactions among large numbers of units, be they areas or points) is the Mathworks® application program MATLAB. The latter is of particular interest for ONS users because of the requirement to handle very large spatial interaction matrices when processing small area data with smoothing methods.

Concluding Comments

The main conclusion about smoothing is that, when faced with data that are highly volatile '*some smoothing of any form is almost always valuable. While the data are changed by smoothing (e.g. from* y_i *to* $\tilde{y}_i$*), one can argue that they are changed only slightly, and that the potential benefits (reduced noise, insight into functional relationships, etc.) far outweigh this concern*' (Kafadar and Horn, 2001:4).

Although the general benefits of smoothing are apparent when faced with noisy data, it is important to remain aware of the limitations of smoothing methods. Critically, we must recognise the subjective nature of the choice of smoothing technology as there are no universal objective reasons for preferring one method over another (Cowling et al., 1996). This is of great consequence for the credibility of the estimates obtained and emphasises the importance of conveying the correct meaning and interpretation of smoothed data to the user.

Given this, the most difficult task for the analyst is to select an appropriate smoothing method given the application context and to specify sensibly how it should operate. Different smoothing techniques may well produce different results when applied to the same data. The larger the degree of smoothing, the smaller the variance of the smoothed value and the larger the bias. We discuss this point in chapter 3 when we consider the theoretical setup for this project. This reduction in variance at the cost of bias in the smoothed value is the fundamental trade-off between bias and variance in all smoothing methods. Marsh (1988) warns that 'this greater freedom brings with it increased responsibility; the choice of how much to smooth will depend on judgement and needs' (p.161).

The outcome of smoothing is linked closely to the choice of the smoothing parameter: a large value induces greater smoothing, highlighting regional patterns; a small value induces less smoothing, emphasising local patterns. The larger the value of the smoothing parameter, the smaller will be the variance of the smoothed value and the larger the bias. This is important, as '[T]he differences between the crude and spatially smoothed estimates are frequently large in geographies with small populations' (Veugelers and Hornibrook, 2002: 102). Similarly, the choice of a weighting function, the area covered and the time span of the input data will determine how individual data points contribute to local smoothed estimates.

We have addressed these issues here by applying a range of smoothing techniques and evaluating their performance within a common framework. This has allowed us to assess the relative performance of different methods and to draw conclusions about the most appropriate methods to use given the results of an empirical evaluation. We set out the details of the methods we implemented and the evaluation framework that we used in the next chapter.

3: Some basic smoothing options

As we saw in chapter 2, smoothing solutions vary widely in their complexity. As this is a preliminary study, we cover only the working and benefits of simple smoothing techniques. Chapter 5 discusses options for further work on the more complex solutions.

This chapter is structured in three parts. While the first part offers an overview of the theoretical fundamentals of smoothing applied to SMRs, the second and third parts concentrate on the specifications used in conducting the analysis reported here, namely the performance criteria that were adopted in the second part and the methods that were applied in the third part.

Fundamentals of Smoothing: Theoretical Setup

The essence of smoothing is that, when data are noisy, an underlying component – free of error – can be made more visible or approximated via smoothing.

While this project fits into the broad view of smoothing, the details of the theoretical setup that underpins our work are defined by the specific aim of this analysis: finding the best estimator of underlying mortality in any particular area. In particular, the approach that we have taken to smoothing in this project rests on a particular theoretical view of the nature of the SMR data under analysis that we set out below.

Each observation y_{it} at every area or ward (*i*) at every moment in time (*t*) has, by definition, two components: mortality proneness or underlying mortality (u_{it}) and random mortality (e_{it}). Formally:

$$\begin{aligned} & y_{it} \equiv u_{it} + e_{it} \\ & e_{it} \sim \left(0, \sigma^2_{e(it)}\right); \\ & E\left(e_{it} u_{it}\right) = 0 \\ & E\left(e_{it} e_{i*t*}\right) = 0 \end{aligned} \qquad (14)$$

Mortality proneness or underlying mortality (u_{it}) is the part in the data linked to the known and unknown determinants of mortality and is due to average rather than individual morbid influences. This is the component that we aim to best retrieve via smoothing.

Random mortality (e_{it}) is that part in the data that cannot be predicted, the noise component[4]. This is assumed to be uncorrelated with the underlying component or

[4] Although the data we are working on are not a sample, there is a useful analogy with sampling that can be made in this case. The random component in (14) is akin to the random component under sampling in that, although its expected value remains unchanged, its specific values vary a lot. In other words, the measurements on the whole population that we have can be viewed as random elements from an infinite super-population.

random components in other areas, past or present. The noise is the part in the data that we aim to reduce or eliminate via smoothing, in order to best uncover the underlying part.

The underlying mortality component in each ward (u_{it}), in turn, has two parts: a structural part and a random part. The structural part (*f*) is the link to past underlying components for that ward, as well as present and past underlying components for its neighbours. The random part $\left(\delta_{it}\right)$ is the part of underlying mortality that remains unexplained by the best prediction based on the structural part and, in that sense, is an intrinsic element of randomness in (un)healthiness and thus in underlying mortality. This component can also be thought of as covering those unmeasured factors that cause systematic differences in mortality between areas in different locations and / or at different moments in time.

This structure can be written as:

$$
\begin{aligned}
u_{it} &\equiv \underset{\substack{i^*,t^*\in S \\ t^*<t \text{ for } i^*=i}}{f}\left(u_{i^*t^*}\right)+\delta_{it} \\
\delta_{it} &\sim \left(0,\sigma^2_{\delta(it)}\right)
\end{aligned} \tag{15}
$$

where *S* is a set of areas (*i**) and points in time (*t**) to which the current value in area *i* is related.

By substituting (15) in (14) it becomes visible that two random elements are confounded in the data, namely δ_{it} and e_{it} :

$$
y_{it} \equiv u_{it} + e_{it} = \overbrace{\underset{\substack{i^*,t^*\in S \\ t^*<t \text{ for } i^*=i}}{f}\left(u_{i^*t^*}\right)}^{\text{structural part}} + \overbrace{\delta_{it} + e_{it}}^{\text{random part}} \tag{16}
$$

The former (δ_{it}) is the random part of the underlying reality which is spatially correlated, while the latter (e_{it}) is noise.

To build an estimate of underlying mortality we need to obtain:

$$
\hat{u}_{it} = \underset{\substack{i^*,t^*\in S \\ t^*<t \text{ for } i^*=i}}{\hat{f}}\left(u_{i^*t^*}\right)+\hat{\delta}_{it} \tag{17}
$$

Firstly, note that the function *f* can be specified in a large number of ways. As the structural part changes from one specification to another, so does the part that remains unexplained - the $\hat{\delta}_{it}$ - which in this sense is a residual or an error term.

Secondly, $u_{i^*t^*}$ are not observed directly. Equation (17) details the components of a smooth estimate, but cannot be implemented as such. In practice, an estimate for the underlying component can be, however, constructed on the basis of the observed values $y_{i^*t^*}$. As equation (18) below shows, these data points have their own *u* and *e* components, as defined by (14), while their *u* components have their own *f* and δ elements, as specified by (15):

$$y_{i*t*} \equiv u_{i*t*} + e_{i*t*} = \underset{\substack{i**,t**\in S \\ t**<t \text{ for } i**=i}}{f} \left(u_{i**t**}\right) + \delta_{i*t*} + e_{i*t*} \quad (18)$$

Therefore, the noise components e_{i*t*} are smoothed over by the application of *f* to the data. Given that $E(e_{ij}) = 0$, smoothing or averaging over the noise leads to its reduction or elimination.

For small area mortality counts, the SMR is a poor estimate of underlying mortality, because the information content of the data is small relative to the random noise (*e*). The goal of smoothing is therefore to increase the information in the data about *u* relative to the random noise, by combining information from other areas and time points. However, such smoothing can only increase the information content about the *f* component of *u*, not the δ component. Only the specific observation in the ward and year of interest can provide information about δ (Best, 2006).

In this context, the role of smoothing is to obtain the best estimate of underlying mortality ($\hat{u}_{it}$) by bringing together the best $\hat{f}$ and the data in such a way that the combination of the two types of error (specification and noise) is minimised: the best *f* is specified and the noise is reduced. For the best specification this occurs while the information contained in the intrinsic random component (δ_{it}) is not lost. In other words, the best smooth is the one that reduces noise as much as possible but retains substantial local differences, like those encapsulated in δ_{it}.

Smoothing is a suitable approach whenever δ_{it} is small relative to e_{it} in (16), an assumption on the basis of which we work. The underlying rationale is that by smoothing the data y_{it}, we smooth (average out) all its components. This leads to a gain as far as reducing noise (e_{it}) is concerned and a loss in relation to the reduction in information (from smoothing δ_{it}). The net result is positive if the gain from smoothing e_{it} more than compensates for the loss from smoothing δ_{it}. This net outcome also depends on the degree of smoothing applied. In other words, there is such a thing as ‘too much smoothing’, in that the loss of information from smoothing can become greater than the gain from noise reduction. This idea can also be understood in terms of variance and bias: a substantial reduction in variance as a result of too much smoothing comes at the cost of introducing a relatively larger bias.

The fact that there is no unique specification of *f* and thus no single solution to the smoothing problem brings forward the need for evaluation to guide the choice of smoothing technique, to which we turn next.

Performance Criteria

All approaches to smoothing can be nested in specification (17) above, as all smoothers are built as a function of values from the past and / or at other locations[5]. Some specifications will be better at retrieving the underlying mortality than others.

In pragmatic terms, it is of interest to assess how much smoothing each technique imposes on the data. When measuring ***smoothness***, we concentrate on the stability of the smoothed data. This can be assessed by comparing, in pairs, outcomes from the application of each technique to data at different points in time. A stable smoother is one for which the difference in the outcomes from one point in time to another is small.

The comparison between any two smoothers is based on the mean squared differences between smooth values across all area units in the study region given repeated applications of the same technique at successive moments in time, t and $(t+1)$ for a particular smoothing technique k:

$$M\hat{S}D = \frac{1}{N}\sum\left(\hat{u}_{t+1,k} - \hat{u}_{t,k}\right)^2 \tag{19}$$

where N is the total number of wards. We report the average $\log\left(M\hat{S}D\right)$, in order to make proportional changes visible and facilitate comparisons.

We can compare any smoothing technique (k) to a chosen benchmark technique (b), by computing the following ratio statistic, where the numerator is the $M\hat{S}D$ for approach k and the denominator is the $M\hat{S}D$ for the benchmark:

$$M\hat{S}DR = \frac{\frac{1}{N}\sum\left(\hat{u}_{t+1,k} - \hat{u}_{t,k}\right)^2}{\frac{1}{N}\sum\left(\hat{u}_{t+1,b} - \hat{u}_{t,b}\right)^2} = \frac{\sum\left(\hat{u}_{t+1,k} - \hat{u}_{t,k}\right)^2}{\sum\left(\hat{u}_{t+1,b} - \hat{u}_{t,b}\right)^2} \tag{20}$$

A value of $M\hat{S}DR$ below 1 for this ratio indicates that option k is more stable than the benchmark one.

Looking back at the underlying ideas of smoothing presented in the theoretical setup above, it becomes clear that a smoother estimate of underlying mortality does not necessarily mean a better estimate. Indeed, it is possible to over-smooth, by either miss-specifying the structural part in (16) or by losing some of the information contain in the δ terms. Indeed, *in extremis*, the national mean is the smoothest estimate for each ward, which does not imply that it also is the best estimate as far as any individual area is concerned. It is important to be able to tell when an over-smooth has been produced.

[5] Estimating underlying mortality is done by borrowing strength in time and space from SMRs only. The use of covariates is possible and could enhance predictive power, but while the outcome of such smoothing would be suitable for targeting resources, it may not be valid for scientific inference. This is because the use of covariate *X* in smoothing *Y* would contaminate the smooth estimate. Analyses of the relationship between *Y* and *X* using the smooth values would be inappropriate because of the confounding influence of *X* which has contributed directly to the smooth value.

We saw in the concluding section of chapter 2 that the choice of smoothing method is not straight forward, as no unique solutions and criteria are available.

Ideally we would evaluate the performance of the various smoothing solutions using cross-validation, a robust form of model evaluation that estimates the generalisation error, i.e. how well the model generalises to new data. The main advantage of cross-validation lies in the fact that evaluation is performed against data that are independent (over space or time) of those used to estimate the model, in a large number of replications. Although we are not able to implement cross-validation in this research for reasons explained below, it is important to clarify what this would entail, in order to understand better the nature of the evaluation framework that we use instead.

In the context of smoothing, the application of cross-validation would mean that smoothed values calculated from some of the data would be compared – across many replications - with independent data for the same areas. The best smoothing method would be the one that correlated best with independent unsmoothed data. This is not to say that the smooth value should be a good estimate of the unsmoothed value, but that a *better* suited smooth would produce values *closer* to the raw data than a less suitable one. Given the design of the evaluation, cross validation would act as a choice mechanism between alternative smoothing solutions. Appendix B provides technical details of this approach.

In order to be able to apply cross-validation, it is necessary to be able to partition the data - repeatedly, a large number of times, across time or across space - into a part used for estimation and a part used for evaluation. Across space, this would be feasible if geo-referenced individual level data were available. In that case, sub-setting the data across space would be practicable. However, in this project we are dealing with area (*lattice)* data, where locations are a discrete set of wards, for which measurements do not vary within the boundaries of each area. Cross validation is therefore not practical.

While independent spatial data are not easily available in the current scenario, independent data in time are available: data used to obtain a smoothed value (in time or space) at moment t are independent from unsmoothed data at moment ($t+1$). As cross-validation requires a large number of replications, its implementation over time is not feasible either with the very short time series that is available to us. However, the availability of some independent data over time means that we can use an evaluation of predictive power as an alternative to cross-validation. In this way the evaluation used takes into account the nature and extent of the data available (see the chapter 4 section "Limitations imposed by the data" for further details).

A smooth series with high ***predictive power*** is one that correlates relatively highly[6] with the un-smoothed data at the next point in time (one step ahead). This acts as a safeguard against over-smoothing, since data that undergoes too much smoothing loses some of its predictive power. The loss of predictive power occurs either because too much smoothing

[6] The actual coefficients of correlation may not be very high in practice, given the presence of noise in the raw data, but what matters here is the relative performance.

reduces the information contained in the random component of mortality δ_{it} (which is contained in the structural part of the smooth value at time *t+1*) or because the function *f* is misspecified. Appendix C provides more detail about this point.

From this perspective, evaluation entails the comparison of the predictive power of more than one predictor on a given outcome. A relatively small difference between actual and predicted values is taken to mean good predictive power; the smaller the difference, the better the smoothing technique. Note that, unlike in evaluation of smoothness, here the predictive power of each smoothing option is to be evaluated against data other than that used to produce the estimate. The measure of validity is the degree of closeness to the actual value at the next point in time.

Using the same notation as above, for two smoothing methods *k* and *m*, the data at any point in time can be decomposed into:

$$\begin{aligned} y_t &= \hat{u}_{t,k} + \hat{e}_{t,k} \\ y_t &= \hat{u}_{t,m} + \hat{e}_{t,m} \end{aligned} \qquad (21)$$

What we are interested in here is the predictive power of the smooth estimates $\hat{u}_{t,k}$ and $\hat{u}_{t,m}$. In particular, we evaluate the correlation between the smooth at one moment in time (*t*) and the rough data one step ahead (*t+1*). Specifically, between any two smoothing solutions *k* and *m* (of which the latter could be a benchmark approach), the one with the lowest mean squared estimated error is to be regarded as the better one. The comparison involves:

$$M\hat{S}E_k = \frac{1}{N}\sum\left(y_{t+1} - \hat{u}_{t,k}\right)^2 \quad vs. \quad M\hat{S}E_m = \frac{1}{N}\sum\left(y_{t+1} - \hat{u}_{t,m}\right)^2 \qquad (22)$$

The pair of estimated mean squared errors can be subject to the ratio comparison described in (20) above.

For both smoothness and predictive power, the strength of the linear relationship between any two series can be measured using Pearson's coefficient of correlation. This is reported alongside the mean squared differences throughout chapter 4.

Some Basic Smoothing Estimators

The smoothing solutions used in this project belong to the broad family of ***linear*** smoothers (see chapter 2 "Data and techniques"). The choice is justified on grounds of relative ease of implementation and of applicability to the type of data under analysis. The application of more elaborate approaches is the main theme of proposed further work (see chapter 5).

In most general terms of this approach, the smooth value of a target variable y_i is defined as:

$$\hat{u}_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}} \qquad (23)$$

where w_{ij} are the weights that link any two observations *i* and *j*, while $\hat{u}_i$ is the smooth value for observation *i*.

The specific approach we take is that of building time and space moving averages over three year time intervals and geographical neighbourhoods defined as the target area and the set of its contiguous neighbours, where the area is the ward.

The smoothing solutions detailed below are specific cases of (23), where values are given to the weights w_{ij} and the identity of the elements *j* (past points or neighbouring values) is specified.

The choice of methods to be implemented was dictated by the nature of the data, the software packages available and the time scale of this project.

Smoothing over time

In the data set we are using, observations are available only for five separate years (see "Limitations imposed by the data" in chapter 4). A time series of five annual observations is very short by the standards of time series analysis and therefore is not suitable for anything other than simple smoothing techniques. Two such techniques – described below - are used in this report and their performance is reported in "Smoothing over time" in chapter 4.

Moving average over three years, with equal weights

The simplest scenario is one where all the values included in the calculation of the smooth have the same weight. Over time, equal weights do not distinguish between the present and the past in terms of the contribution to the smooth estimate, which is specified as:

$$\hat{u}_t = \frac{1}{T}\sum_{k=0}^{T-1} y_{t-k} \tag{24}$$

where the subscript *t* refers to the year to which the smoothing is applied and T denotes the length of the time-horizon for smoothing.

Moving average over three years, with declining weights

Declining weights assign more importance to recent data points and less to points further in the past. We specify two sets of weights, corresponding to high persistence or memory (relatively high weights for the past; subscript *h*) and low persistence or memory (relatively low weights for the past, subscript *l*). For the generic form:

$$\begin{aligned} \hat{u}_t &= \sum_{k=0}^{T-1} \alpha_k y_{t-k} \\ \sum_{k=0}^{T-1} \alpha_k &= 1 \end{aligned} \tag{25}$$

we set *T=3* and apply the following weighting schemes:

$$\begin{aligned} &(i) \quad \alpha_{0h} = 0.5; \alpha_{1h} = 0.3; \alpha_{2h} = 0.2 \quad \text{(high memory)} \\ &(ii) \quad \alpha_{0l} = 0.7; \alpha_{1l} = 0.2; \alpha_{2l} = 0.1 \quad \text{(low memory)} \end{aligned} \tag{26}$$

Smoothing in space

In spatial analysis terminology, ward data belongs to the category of ***lattice data***, i.e. observations from a random process over a countable collection of contiguous spatial units of varying size (the wards), together with a *neighbourhood structure*. In the ***neighbourhood*** structure used here, the neighbours are defined as wards that border on each other. This structure means that, for each individual ward, the areas that matter in terms of spatial interaction are the immediate neighbours.

By analogy with smoothing over time, the simplest way of smoothing in space is by replacing each individual value (for the target area) by a weighted average of itself and its neighbours. In our study we consider neighbours defined by first order (immediate) and second order (neighbours of first order neighbours) contiguity[7] and with two types of *ad hoc* weights: equal and tapering over space, to mirror the structure of the time smoothing solutions described above .

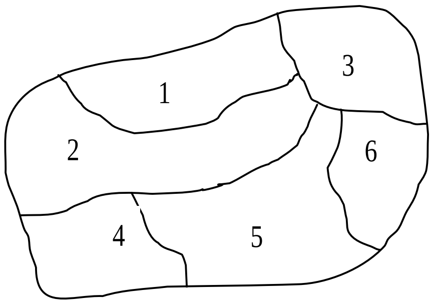
In general, neighbourhoods based on contiguity are specified through a matrix of spatial interactions ***W***, whose elements are defined like in equation (4), with neighbours identified through their sharing a common border:

$$w_{ij} = \begin{cases} 1 & i \text{ and } j \text{ share common border} \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

This approach is illustrated in Figure 3.1, which contains the map of a set of areas and the corresponding spatial interactions matrix using first-order neighbours. In this matrix, a value of 1 indicates that two areas are adjacent to one another, while a 0 indicates that the areas are not adjacent. The first line of the matrix contains information about the identity of the neighbours for area 1 (who shares a border with areas 2 and 3, the only non-zero values in the row). The second line contains the relevant information for area 2, and so on.

[7] Spatial interaction among neighbours can also be expressed as a function of distance between the centroids of the areas involved, as a function of shared border or as a function of some flow variable (e.g. migration). Also, the weights can be specified *ad hoc* or defined through the use of a kernel. See section 5.3 for further details.

Figure 3.1 Simple contiguity setup



$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

The idea can be extended to second-order neighbours, which for any area *i* means the contiguous (first order) neighbours of its first order neighbours, i.e. its neighbours once removed.

In this respect, we use three types of specifications: with first order neighbours only, with first and second order neighbours taken together and with weights that differentiate between first and second order neighbours, as explained below.

In this report we work solely with this fundamental binary framework, as it is the most commonly used. Alternative specifications, like those presented in the "Data and techniques" section of chapter 2, may be explored in a future project on this subject.

Spatially weighted average over contiguous areas with equal weights

The spatial equivalent of the time moving average with equal weights (24) is the equal weights average of the values for each target area and its specified neighbours:

$$\hat{u}_i = \frac{1}{n_i + 1}\left(y_i + \sum_{j \neq i}^{n_i} y_j \right) \tag{28}$$

Here the subscripts *i* and *j* refer to the ward to which the smoothing is applied and, respectively, its neighbours, while n_i is the number of neighbours of that ward. We have applied this approach with two sets of neighbours for each area: first order neighbours only, and first and second order neighbours taken together.

Spatially weighted average over contiguous areas, with unequal weights

Similarly, the spatial counterpart of the time decay formulation (25) of the smooth is a spatially weighted average that distinguishes between the target area and its neighbours, through a spatial decay parameter.

This approach is applied separately for the inclusion of first-order and first- and second order neighbours in the spatial moving average. Specifically, three values for the parameter (α) are used in the specification using only first order neighbours: a high, a medium and a low value. These vary the degree of importance allocated to the target area in the production of the smooth estimate and thus the relevance of the neighbourhood in this computation. A high value of the decay parameter means that a higher weight is placed on the target area than on its neighbours and thus represents a low spatial link situation. The converse is true for a relatively low value of the decay parameter.

$$\hat{u}_i = \alpha y_i + (1-\alpha)\frac{\sum_{j \neq i}^{n_i} y_j}{n_i}, \quad n_i \neq 0 \qquad (29)$$
$$\alpha = 0.9; \quad \alpha = 0.7; \quad \alpha = 0.5;$$

Once more, the subscripts *i* and *j* refer to the ward to which the smoothing is applied and, respectively, its neighbours, while n_i is the number of contiguous neighbours of that ward.

Note that, for the case of zero neighbours (one occurrence in the data set used here), no pure spatial smoothing is applied, i.e. $\alpha = 1$:

$$\hat{u}_i = y_i, \quad n_i = 0 \qquad (30)$$

We also investigate the performance of the spatial weighted average that distinguishes between the influence of immediate neighbours and neighbours of neighbours (neighbours once removed) through the use of separate matrices of spatial interaction and separate smoothing weights:

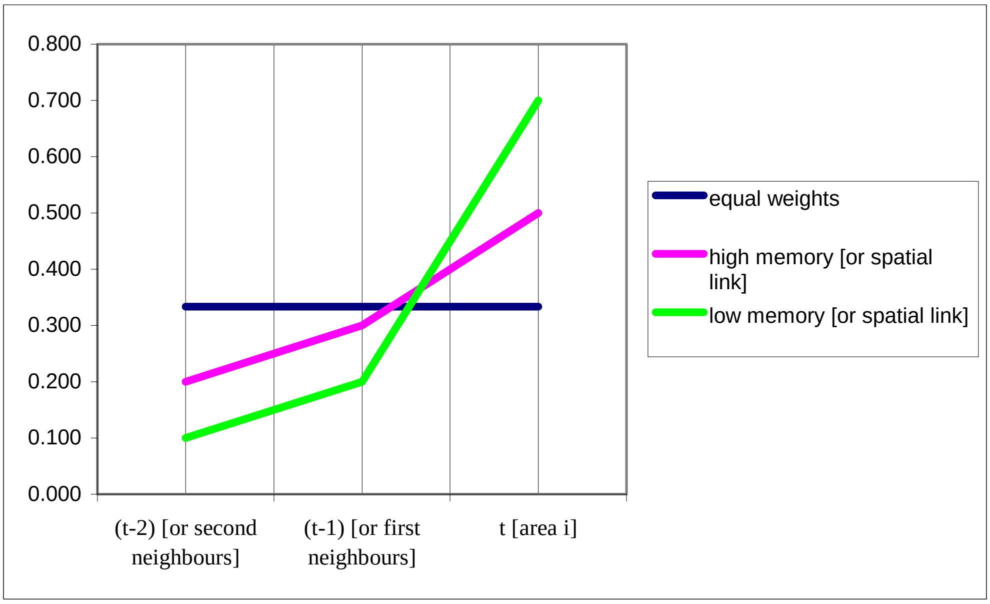
$$\hat{u}_i = \sum_{k=0}^{2} \alpha_k L_k(y_i) \quad \text{and} \quad \sum_{k=0}^{2} \alpha_k = 1$$
$$L_k(y_i) = W_k y_i \qquad (31)$$
$$W_0 = I$$

where L is the spatial lag operator, W_k a matrix of spatial interactions of order k and I an identity matrix. This mirrors the structure of the time-smooth with decaying memory in equation (25) in the report.

The “Smoothing in space” part of Chapter 4 reports the results of spatial smoothing using these approaches.

The three weighting schemes used for the time smoothing and for the spatial smoothing with first and second order neighbours are depicted in Figure 3.2.

Figure 3.2: Smoothing weights across time [or space]



Smoothing over time and space

Smoothing in both time and space can be achieved by either smoothing in time the spatially smooth data or by smoothing in space the data smoothed over time.

In our analysis, we evaluate the performance of smoothers in time and in space separately. The best performing candidates of each type are then combined to produce estimates that are smoothed across both time and space. The "Smoothing over time and space" section of chapter 4 reports the results of the specific combination that was implemented. In Appendix D, we set out the relevant technical details of these combined approaches.

4: Results

The performance of the proposed smoothing solutions when applied to ward-level SMR data is reported below. The section is structured in four parts: the first one discusses the limitations imposed by the data available for analysis, while the remaining three report on the performance of time smoothing, spatial smoothing and smoothing over time and space, respectively. The stability of the smooth values and their predictive power are both covered. For the best candidate method in each category we also report on their performance on subsets of the data (i.e. urban vs. rural wards) and when we smooth death and population counts separately and then recombine them to form an SMR.

Chapter 5 and Appendix E contain maps of the raw data and best smooth in the analysis. Appendix F contains an application of the best smooth to the *status quo* five-year SMRs.

All the variables used and produced in this analysis are listed in Appendix G. The MATLAB functions for the implementation of time and space smoothing are given in Appendices H and I, respectively[8].

Limitations Imposed by the Data

Mortality Data

The deaths and population data made available for this analysis are defined across 8797 Census Standard Table wards, with deaths assigned to wards using the May 2005 National Statistics Postcode Directory.

The ward level SMR data published currently by the ONS were generated through basic smoothing in time by aggregating deaths and population data over the five years 1999 - 2003. As there is only one such series available to date, no direct evaluation of its performance is possible. Instead, we construct a ***benchmark case*** whose performance can be evaluated and against which the proposed alternatives can be assessed. Thus, SMRs built on totals over three (rather than five) years were made available alongside the five-year aggregated measure and the annual data from which these aggregates were constructed (see Appendix A for details of calculation). The outcome is three benchmark smooth series, corresponding to the intervals 1999-2001, 2000-2002, and 2001-2003.

In establishing the benchmark case one is confronted by data problems and uncertainty in calculating SMRs, given the need to estimate small-area population for the inter-censual years and the risk of large errors. A previous version of this report covered the findings of an evaluation of smoothing SMRs calculated using 2001 Census population measures. Here we report on the evaluation of smoothing SMRs calculated using 2001 ward-level population estimates. Note that the two sets of results are not directly comparable, as they differ not only in the population measure used in calculating the SMRs, but also in terms

[8] No separate script is needed for space-time smoothing, as this can be achieved by applying one smoothing script to the data smoothed using the other.

of geography and specifics of computation. Is it important to note, however, that for both data sets, the relative performance of the solutions is unchanged.

As discussed in the “Performance criteria” part of chapter 3, evaluation through cross-validation is not possible here since data are only available at ward level. Cross-validation would have been possible with access to individual geo-referenced data. Consequently, we use the predictive power indicator as an alternative measurement of suitability for candidate smoothing techniques.

The data available impose limitations on the techniques applicable and the evaluation setup. They also constrain the conclusions and recommendations that can be meaningfully formulated. The performance of the options explored can only be assessed against that of the raw data and of a benchmark case constructed for this purpose on the basis of data for three rather than five years. The findings remain relevant to the five-year SMRs only inasmuch as their three-year counterparts act as good proxies. While it is reasonable to assume that this will be the case, further work on a time series of five-year SMRs would be necessary in order to confirm or reinforce their validity.

Spatial Heterogeneity

An assumption often made when analysing spatial data is that of spatial homogeneity, i.e. the absence of variation in a process under analysis with regards to the location of the measurement.

As far as SMRs are concerned, the stability of the measure depends on the size of the population at risk on which it is based. In the case of wards, populations vary considerably (from less than 1000 to over 30,000). As a result, measures of mortality for highly populated areas can be expected to be more stable than those for sparsely populated areas. In our case, differences in the stability of SMRs between urban and rural areas can be expected.

In this preliminary study we start from the assumption of spatial homogeneity, as the smoothing techniques that we evaluate are simple ones and therefore not equipped to differentiate among types of areas. While basing our choice of smoothing specification on criteria using overall performance, we monitor the performance of our chosen smoothing solutions in terms of spatial heterogeneity and use our findings on the differences between urban and rural performance and the separate smoothing of numerators and denominators to inform the proposed refinements to be implemented in further work on this topic (see chapter 5).

Geographical Data

This is the first study in ONS to utilise digital geographical boundaries as an input to mathematical models. Specifically, we have used digital ward boundaries to extract information about neighbouring areas, and this information has been used as a key input to the spatial smoothing methods described in chapter 3 and reported on here. In fact, there are two digital ward geographies available for 2003: the Extent of the Realm (EOR) geography and the clipped (CP) geography. The EOR boundary sets include the foreshore of coastal areas, which in some cases (e.g. The Wash, Morecambe Bay) may amount to many square miles of uninhabited mud. The boundaries of administrative areas such as wards are extended out to the low water mark. The CP boundaries do not include the foreshore. Their coastal limits are drawn along the mean high water level.

Each of these geographies presents problems for the establishment of spatial contiguity. In general, the EOR geography tends to overstate the level of contiguity, while the CP geography tends to understate it. While the global differences in performance that arise are small[9] , it is likely that specific areas will experience particular local effects as a result of choosing a particular digital boundary set for analysis. We have not considered this issue here because of time constraints, but it does impact on analysis outcomes and represents an important factor that will require careful thought as the use of spatial analysis methods becomes more prevalent.

In the current version of the report we only present findings based on the EOR geography.

Smoothing Over Time

Smoothing over time covers intervals of three years. Thus, from annual data between 1999 and 2003 three smooth values can be obtained for each ward: 1999 to 2001, 2000 to 2002 and 2001 to 2003.

Table 4.1 gives the correspondence between the names of the time smoothing methods referred to and the technical content thus abbreviated.

Table 4.1 Summary of time smoothing methods

Name Used	Content
Status quo	Three-year moving average with equal weights (eq.24)
High memory	Three year moving average with slowly declining weights (eqs.25 & 26i)
Low memory	Three year moving average with quickly declining weights (eqs. 25 & 26ii)

Smoothing performance

Table 4.2 summarises the performance of the status quo and the proposed solutions in terms of stability of the smooth values. The (log) mean squared differences (in level and as ratio to the benchmark value, according to (19) and (20)) are reported alongside the coefficient of correlation. The variance of the raw and smoothed values across all areas is also reported.

The first five lines describe the annual raw data, in terms of variance and stability. What becomes immediately visible is that time smoothing increases the stability in the data, from around 0.58 (in the raw data) to between around 0.79 (low memory moving average case) and 0.93 (equal weights moving average case), i.e. by between 36 per cent and 60 per cent.

[9] The differences between the values reported previously were no more than 0.06 for the correlation coefficient and no more than 0.02 for the log mean squared difference.

Table 4.2: Stability of smoothing over time

Method	Time scale	Variance	Performance		
			Correlation coefficient	LMSD*	LMSDR**
Raw data	1999	894	0.592		
	2000	783.6	0.575		
	2001	763.3	0.567		
	2002	771.9	0.568		
	2003	790.3			
status quo	1999/2001	585.7	0.928	4.44	1
	2000/2002	550.8	0.928	4.37	1
	2001/2003	549.4			
high memory	1999/2001	585	0.897	4.79	1.08
	2000/2002	564.8	0.894	4.79	1.09
	2001/2003	569			
low memory	1999/2001	625.1	0.794	5.55	1.25
	2000/2002	617.8	0.79	5.57	1.27
	2001/2003	627.6			

* Log mean squared differences; ** Log mean squared difference ratio to the *status quo* value.

Of the moving average smoothers explored, the most stable is the one with equal weights, i.e. *status quo*. The stability of the time smoothed data is lower in the case of the moving average with a tapering effect and the reduction in stability is more visible in the specification with the lowest memory. For the low memory case of the moving average the coefficient of correlation one step ahead is just over 0.79.

The overall pattern is not surprising, since over the time horizon under analysis there are no reasons to expect that the past matters any less than the present, as far as identifying the underlying component is concerned. The underlying process can be expected to change very slowly, so the degree of inertia is high. Note that the least stable is the smoother that gives the highest weight to the current values. In the extreme, such smoothing is equivalent to moving back to the un-smooth series.

Evaluation through prediction

The predictive power of the smoothed series obtained through the application of various smoothing techniques can be used to identify the most suitable technique for the given data set, i.e. the one that produces the smooth data with the highest predictive power. The differences in outcome for the various time-smoothing solution explored are fairly small. However, it is not the magnitude but the direction of change in predictive power that matters here.

The predictive power of the time smoothed data is summarised in Table 4.3. The relative positions in terms of predictive performance mirror the ones based on stability described above and are as follows: the best performer is the time smooth based on a three-year moving average with equal weights (the status quo, rows highlighted in the table). Time-smoothing with a weighted moving average has the lowest predictive power, and this is lower in the case of low memory than in the case of high memory.

On this basis, it is the moving average with equal weights (24) that we select as the time smoothing solution to combine with the best performing spatial smoothing solution in order to produce space-time smoothing (see “Smoothing over time and space” below).

Table 4.3: Predictive power of time-smoothed data

Method	Time scale	Performance		
		Correlation coefficient	LMSD*	LMSDR**
status quo	1999/2001	0.662	6.15	1
	2000/2002	0.659	6.16	1
high memory	1999/2001	0.658	6.16	1
	2000/2002	0.654	6.18	1
low memory	1999/2001	0.632	6.25	1.02
	2000/2002	0.63	6.27	1.02

* Log mean squared differences; ** Log mean squared difference ratio to the *status quo* value.

Smoothing in Space

Smoothing in space is applied separately to each set of annual data between 1999 and 2003. Specification of the neighbourhoods that define the spatial interactions in the data is based on the EOR geography. Table 4.4 gives the correspondence between the names of the spatial smoothing methods referred to and the technical content thus abbreviated.

Table 4.4: Summary of space smoothing methods

Name Used	Content
one lag, equal weights	equal weights spatial moving average using first order neighbours only (eq.28)
one lag, high link	unequal weights spatial moving average using first order neighbours only, equal weights between each area and the average value for its neighbourhood (eq.29, a=0.5)
one lag, medium link	unequal weights spatial moving average using first order neighbours only, slightly higher weights for each area compared to the weights for its neighbourhood (eq.29, a=0.7)
one lag, low link	unequal weights spatial moving average using first order neighbours only, higher weights for each area compared to the weights for its neighbourhood (eq.29, a=0.9)
one lag, equal weights, enlarged	equal weights spatial moving average using first and second order neighbours (eq.28)
one lag, high link, enlarged	unequal weights spatial moving average using first and second order neighbours, equal weights between each area and the average value for its neighbourhood (eq.29, a=0.5)
one lag, medium link, enlarged	unequal weights spatial moving average using first and second order neighbours, slightly higher weights for each area compared to the weights for its neighbourhood (eq.29, a=0.7)
one lag, low link, enlarged	unequal weights spatial moving average using first and second order neighbours, higher weights for each area compared to the weights for its neighbourhood (eq.29, a=0.9)
two lags, high link	spatial moving average using weights that differentiate between first and second order neighbours, with slightly higher weight for the first order neighbours relative to the second order ones (eq.31 & 26i)
two lags, low link	spatial moving average using weights that differentiate between first and second order neighbours, with higher weight for the first order neighbours relative to the second order ones (eq.31 & 26ii)

Smoothing Performance

The number of neighbours per ward ranges from 0 (for one ward) to 19 first order neighbours and 43 second order neighbours. The average number of first order neighbours per ward is 5.86 and 14.32 second order neighbours, while the median is 6 for first order neighbours and 14 for second order ones. Table 4.5 contains the two measures of stability (coefficient of correlation and mean squared differences), for every pair of annual series. They correspond to one-step ahead (one year, in this case) stability.

Table 4.5: Stability of smoothing over space

Method	Years	Correlation coefficient	LMSD*	LMSDR**
one lag, equal weights	1999-2000	0.822	4.8	
	2000-2001	0.813	4.7	
	2001-2002	0.806	4.69	
	2002-2003	0.801	4.71	
	average	0.811	4.72	1.07
one lag, high link	1999-2000	0.75	5.38	
	2000-2001	0.737	5.31	
	2001-2002	0.732	5.3	
	2002-2003	0.726	5.32	
	average	0.736	5.33	1.21
one lag, medium link	1999-2000	0.679	5.88	
	2000-2001	0.663	5.83	
	2001-2002	0.657	5.82	
	2002-2003	0.654	5.84	
	average	0.663	5.84	1.33
one lag, low link	1999-2000	0.618	6.34	
	2000-2001	0.601	6.29	
	2001-2002	0.594	6.29	
	2002-2003	0.593	6.31	
	average	0.601	6.31	1.43
one lag, equal weights, enlarged	1999-2000	0.903	3.92	
	2000-2001	0.903	3.68	
	2001-2002	0.896	3.65	
	2002-2003	0.894	3.65	
	average	0.899	3.73	0.85
one lag, high link, enlarged	1999-2000	0.742	5.2671	
	2000-2001	0.734	5.1763	
	2001-2002	0.726	5.1783	
	2002-2003	0.721	5.1977	
	average	0.731	5.2048	1.18
one lag, medium link, enlarged	1999-2000	0.665	5.8622	
	2000-2001	0.653	5.795	
	2001-2002	0.645	5.7993	
	2002-2003	0.642	5.8176	
	average	0.651	5.8185	1.32

one lag, low link, enlarged	1999-2000	0.612	6.3403	
	2000-2001	0.597	6.2843	
	2001-2002	0.589	6.2893	
	2002-2003	0.588	6.3067	
	average	0.597	6.3051	1.43
two lags, high link	1999-2000	0.748	5.2894	
	2000-2001	0.738	5.2066	
	2001-2002	0.731	5.2019	
	2002-2003	0.726	5.2218	
	average	0.736	5.2299	1.19
two lags, low link	1999-2000	0.672	5.8685	
	2000-2001	0.659	5.8056	
	2001-2002	0.651	5.8056	
	2002-2003	0.648	5.8241	
	average	0.658	5.826	1.32

* Log mean squared differences; ** Log mean squared difference ratio to the *status quo* value.

The gain in stability from spatial smoothing ranges between 4 per cent and 56 per cent; the coefficient of correlation one step ahead varies between around 0.60 (one lag, low link, enlarged) and 0.89 (one lag, equal weights, enlarged), compared to around 0.58 in the raw data (see Table 4.2).

Table 4.6: Variances in the raw and spatially smoothed data (five year average)

Method	Variance
raw data	800.6
one lag, equal weights	291.5
one lag, high link	387.8
one lag, medium link	513.7
one lag, low link	692.5
one lag, equal weights, enlarged	189.5
one lag, high link, enlarged	335.9
one lag, medium link, enlarged	483.7
one lag, low link, enlarged	682.8
two lags, high link	351.4
two lags, low link	496.2

The largest gain in stability comes from spatial smoothing with equal weights, where first- and second-order neighbours are included in a single lag (one lag, equal weights, enlarged). This is the solution that brings the largest reduction in variance of the data, as Table 4.6 shows. In extremis, the most stable smooth would be the one that replaced the value for every ward by the mean across all wards in the area considered. That would also be the solution that introduced most bias at local level. This trade-off between variance and bias is not visible when the data is smoothed over time, hence the consistency in rankings between stability and predictive power reported in section 4.2. The picture in the case of smoothing across space is different, as the results in section 4.3.2 clarify.

Evaluation through prediction

As indicated by both the mean squared differences and the correlation coefficients reported in Table 4.7, the spatial smoothing solution with strongest predictive power is the spatial weighted average that differentiates between first- and second-order neighbours, with high spatial link profile (highlighted rows in the table).

Compared to the spatially weighted average with equal weights (one lag) that, as we have seen in part 4.3.1, produces the most stable smooth values, the solution with the highest predictive power preserves more of the variance and thus introduces less bias (see Table 4.6). We favour predictive power over stability for this reason.

On the basis of the predictive power, it is the spatially weighted average with two lags, with unequal weights (31), the high spatial link scenario (26i) that we select as the space smoothing solution to combine with the best performing time smoothing solution selected above in order to produce space-time smoothing.

Table 4.7: Predictive power of spatially smoothed data

Method	Years	Correlation coefficient	LMSD*	LMSDR**
one lag, equal weights	1999-2000	0.542	6.35	
	2000-2001	0.546	6.29	
	2001-2002	0.526	6.33	
	2002-2003	0.521	6.36	
	average	0.534	6.34	1.44
one lag, high link	1999-2000	0.622	6.22	
	2000-2001	0.617	6.18	
	2001-2002	0.602	6.21	
	2002-2003	0.602	6.23	
	average	0.611	6.21	1.41
one lag, medium link	1999-2000	0.62	6.29	
	2000-2001	0.609	6.24	
	2001-2002	0.597	6.27	
	2002-2003	0.598	6.29	
	average	0.606	6.27	1.42
one lag, low link	1999-2000	0.603	6.44	
	2000-2001	0.588	6.4	
	2001-2002	0.578	6.41	
	2002-2003	0.579	6.43	
	average	0.587	6.42	1.46
one lag, equal weights, enlarged	1999-2000	0.481	6.41	
	2000-2001	0.499	6.35	
	2001-2002	0.479	6.39	

	2002-2003	0.473	6.42	
	average	0.483	6.4	1.45
one lag, high link, enlarged	1999-2000	0.621	6.2016	
	2000-2001	0.619	6.1614	
	2001-2002	0.602	6.2038	
	2002-2003	0.603	6.2239	
	average	0.611	6.1977	1.41
one lag, medium link, enlarged	1999-2000	0.615	6.28	
	2000-2001	0.606	6.24	
	2001-2002	0.593	6.27	
	2002-2003	0.594	6.28	
	average	0.602	6.27	1.42
one lag, low link, enlarged	1999-2000	0.601	6.44	
	2000-2001	0.586	6.39	
	2001-2002	0.576	6.41	
	2002-2003	0.577	6.43	
	average	0.585	6.42	1.47
two lags, high link	1999-2000	0.626	6.2	
	2000-2001	0.622	6.16	
	2001-2002	0.606	6.2	
	2002-2003	0.608	6.22	
	average	0.616	6.19	1.41
two lags, low link	1999-2000	0.619	6.28	
	2000-2001	0.609	6.24	
	2001-2002	0.597	6.26	
	2002-2003	0.598	6.28	
	average	0.606	6.26	1.42

Smoothing over time and space

The best performing smoothing solution across time was the moving average with equal weights (24). The best smoothing option over space was the spatially weighted average over first- and second-order neighbours (two spatial lags) with a high spatial link (31 & 26i). These two smoothers are the ones that are combined to obtain space-time smoothing. As anticipated in section 3.3.3, smoothing in time and space can be achieved by either smoothing the spatially smoothed data across time or by smoothing the data smoothed over time across space. This amounts to applying (24) to values calculated using (31) or vice versa. However, only one of the two needs to be implemented, as for these specifications smoothing the spatially smoothed data over time and spatially smoothing the time smoothed data are equivalent (see Appendix D for proof).

Smoothing performance

The stability of data smoothed over time and space is reported in Table 4.8, which contains the results of spatial smoothing applied to the data smoothed over time using the status quo method. The combination of smoothing over time with smoothing over space produces the most substantial gains in stability compared to the raw data, with an increase of around 67 per cent.

Table 4.8: Stability of smoothing over time and space

Method	Estimates		Performance	
	Years		Correl. coef.	log(MSD*)
two-lags, high link, spatial smoothing of *status quo*	1999/2001	2000/2002	0.961	3.23
	2000/2002	2001/2003	0.961	3.08

* Mean Squared Difference

Evaluation through prediction

The predictive power of data smoothed over time and space is reported in Table 4.9. The gain from space-time smoothing is unequivocal, as the predictive power of the space-time smooth is higher than that of the raw data by around 15 per cent.

Table 4.9: Predictive power of data smoothed over time and space

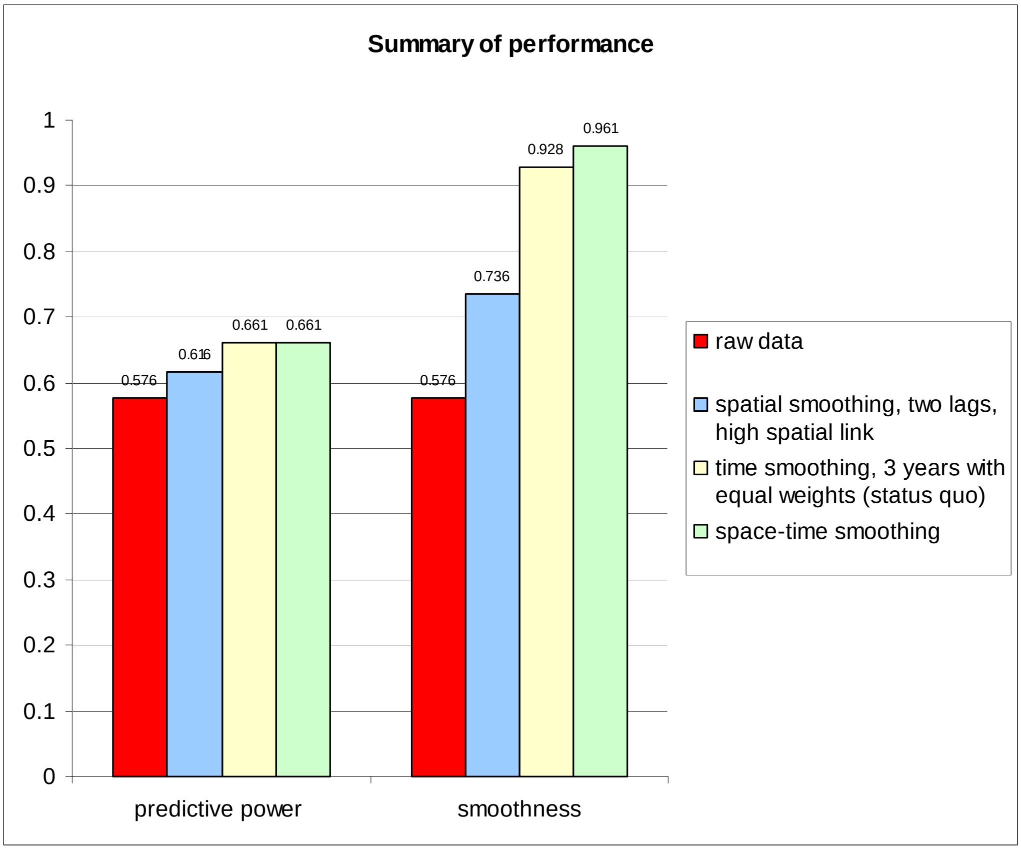
Method	Estimates	Raw data	Performance	
			Correl. coef.	log(MSD*)
two-lags, high link, spatial smoothing of status quo	1999/2001	2002	0.662	6.08
	2000/2002	2003	0.661	6.11

* Mean Squared Difference

Graphically, the overall situation is summarised in Figure 4.1., which depicts the relative performance of the best smoothing solutions identified against that of the raw data, which is set as the base (equal to 1). The gain from smoothing in terms of both stability and predictive power is clearly visible for pure time, pure space and space-time smoothing. The largest increase in stability comes from the application of space-time smoothing. When considering the chart, the reader should keep in mind that, as far as predictive power is concerned it is the relative positions of the columns that matters, not the specific magnitudes.

Reassuringly, the status quo method has the best performance amongst the time smoothing solutions that were evaluated. Further gains in stability and predictive power can be obtained through spatially smoothing the time-smoothed values.

Figure 4.1: Gains in predictive power and stability achieved by smoothing .



Comparing performance in rural and urban areas

An important problem when mapping SMRs by administrative areas is that the rates are estimated with different precisions, depending on the population size in each area. The methods explored here so far – including the best-performing one - provide estimates that have less variance than the raw SMRs, but still exhibit differences between rural and urban areas. These differences undermine the interpretation of patterns across space (on a map), because estimates of different precision are being compared.

An exploration of the extent of this problem produced the results reported in tables 4.10 and 4.11. Here, we compare results across rural and urban definitions from the National Statistics Urban-Rural classification. Both the mean and the variance of the raw and smooth values are systematically lower in the case of rural areas.

Table 4.10: Summary statistics for urban and rural areas.

Variable	Mean		Variance	
	Urban	Rural	Urban	Rural
Raw data				
SMR99	104.6	88.5	870.9	772.2
SMR00	101	86.5	766.5	679.9
SMR01	99.2	83.9	737.3	670.9
SMR02	99.5	85	734.8	735.3
SMR03	100.1	85.6	757.4	740.9
Time smoothing: equal weights (status quo)				
SMR3E9901	101.6	86.3	592.5	382.3
SMR3E0002	99.9	85.1	554.6	370
SMR3E0103	99.6	84.8	549	382.9
Spatial smoothing: two lags, high link				
SMR99SpH	103.7	89.8	398.2	252.2
SMR00SpH	100.2	87.8	342.7	216.5
SMR01SpH	98.3	85.4	339.4	220.6
SMR02SpH	98.6	86.3	327.9	232.2
SMR03SpH	99.2	86.9	338.7	238.4
Space-time smoothing: two-lag, high link, spatial smooth of status quo				
SMR9901SpH	100.7	87.6	305.6	141.7
SMR0002SpH	99	86.5	283.9	135.2
SMR0103SpH	98.7	86.2	281.3	141.9

The results for all cases of interest – raw values and best performing method in each category – indicate that systematic differences in performance are present. As a starting point, the raw data for urban areas is much more stable and has almost twice the predictive power than that for rural areas.

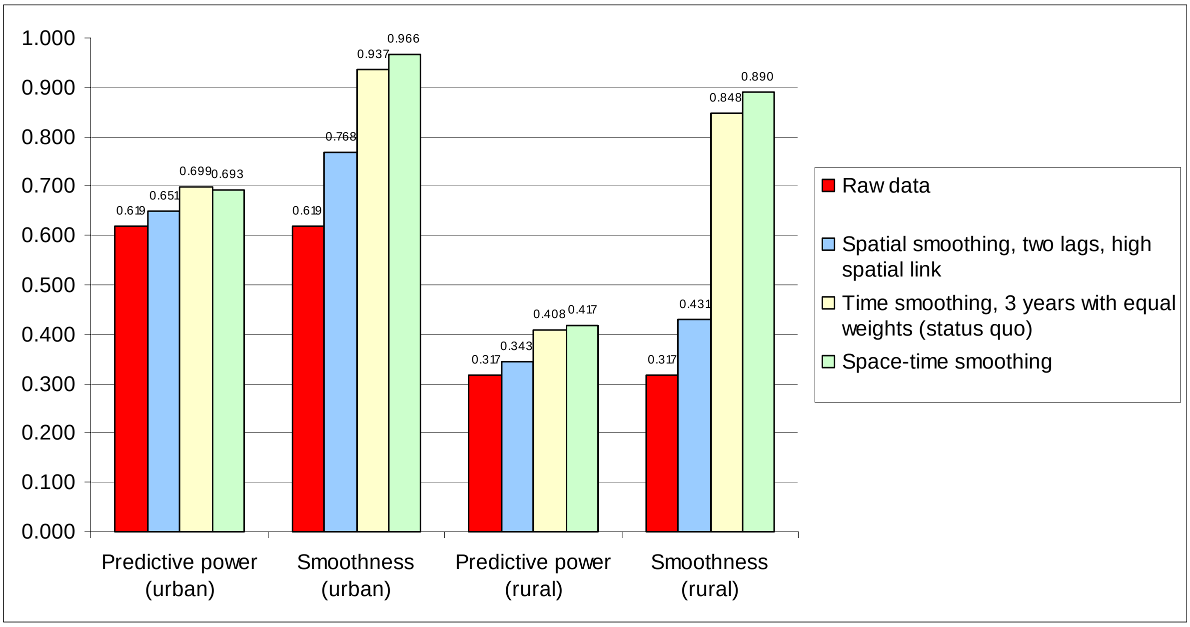
Table 4.11: Subset performance in urban and rural areas.

Variable	Correlation Coefficient			Log Mean Sq. Difference		
SMOOTHNESS (one year apart)	Overall	Urban	Rural	Overall	Urban	Rural
Raw data	0.576 (0.0001)	0.619 (0.0002)	0.317 (0.0006)			
Time smoothing: equal weights(status quo)	0.928 (0.0002)	0.937 (0.0000)	0.848 (0.0000)	4.41 (0.0022)	4.28 (0.0020)	4.75 (0.0026)
Spatial smoothing: two lags, high link	0.736 (0.0001)	0.768 (-0.0001)	0.431 (0.0005)	5.23 (0.0016)	5.1 (-0.0029)	5.57 (0.0012)
Space-time smoothing: two lag, high link, spatial smooth of equal weights time smooth (status quo)	0.961 (0.0000)	0.966 (0.0000)	0.890 (0.0000)	3.16 (0.0115)	3.05 (0.0139)	3.44 (0.0067)
PREDICTIVE POWER (one year ahead)						
Raw data	0.576 (0.0001)	0.619 (0.0002)	0.317 (0.0006)			
Time smoothing: equal weights(status quo)	0.661 (0.0001)	0.699 (0.0000)	0.408 (0.0003)	6.16 (0.0000)	6.01 (0.0003)	6.53 (0.0008)
Spatial smoothing: two lags, high link	0.616 (0.0001)	0.651 (0.0001)	0.343 (0.0005)	6.191 (0.0006)	6.08 (-0.0009)	6.50 (0.0017)
Space-time smoothing: two lag, high link, spatial smooth of equal weights time smooth (status quo)	0.661 (0.0000)	0.693 (0.0000)	0.417 (0.0005)	6.094 (0.0004)	5.97 (0.0015)	6.42 (0.0005)

Note: Figures in brackets are variances over time.

Details of the performance on urban and rural subsets appear in Figure 4.2. It is important to note that the relative positions in terms of performance for the methods compared are the same for the two subsets of data. The stability (smoothness) of the urban and rural values increases to comparable levels. The increase in predictive power due to smoothing is also present in both cases. In the case of urban values, the gain in predictive power from time smoothing is marginally higher than that from space-time smoothing (0.6990 vs. 0.6925). This is not the case overall or for the rural areas. This may be regarded as an indication that a spatial smoothing solution that differentiates between urban and rural areas in terms spatial structure and interactions would be more suitable.

Figure 4.2: Relative performance of smoothing methods in urban and rural areas



Accounting for population differences by separate smoothing of SMR numerator and denominator

We have seen that the variability of ward-level populations results in geographical variation in smoothness and predictive power. One way to take account of this issue (Kafadar, 1996) might be to smooth the death count (numerator) and population count (denominator) separately and then to compute the SMR from the resulting smoothed values. In this section, we compare the results of smoothing the numerator and denominator separately to the best results obtained from smoothing the SMR as a single quantity as detailed above.

The best performing space smooth when the SMR is treated as a single quantity is a spatial moving average using weights that differentiate between first and second order neighbours, with slightly higher weight for the first order neighbours relative to the second order ones. The best time smooth for the single quantity SMR is a three year moving average with equal weights. The best space-time smooth for the single quantity SMR is a combination of the best time smooth and the best space smooth.

In this section, we compare the results from these smoothers with equivalents computed by smoothing the numerator and denominator of the SMR separately. The estimators deployed are identical, save for the fact that two values are computed instead of one, so the time smoothed SMR generated by smoothing numerator and denominator separately is a three year moving average with equal weights, the corresponding space smooth is a moving average with unequal weights that differentiate between first and second order neighbours and the space-time smooth is the same optimal combination as that deployed for single values.

Tables 4.12 to 4.17 show the log mean squared differences (LMSD) between the estimator and the observed value in the reference year, the log mean square errors (LMSE) in the reference year and the ratios between separate smoothing of numerator and denominator and our previous results from smoothing the SMR as a single value. The correlation coefficient shown in the tables is a Pearson Product-Moment Correlation Coefficient which measures the strength of the association between the smooth values obtained using separate numerator and denominator values and smoothing using the single SMR value.

Table 4.12: Stability of space smoother where numerator and denominator are processed separately compared with space smoother derived from single SMR value.

Method	Timescale	Correlation coefficient (Separate SMR denominator and numerator vs. single SMR)	LMSD*	LMSDR** (Separate SMR denominator and numerator vs. single SMR)
Best space smooth when SMR numerator and denominator are treated separately	1999 – 2000	0.9830	5.13	0.97
	2000 – 2001	0.9795	5.05	0.97
	2001 – 2002	0.9800	5.05	0.97
	2002 – 2003	0.9805	5.07	0.97
Best space smooth from single SMR value	1999 – 2000		5.29	
	2000 – 2001		5.21	
	2001 – 2002		5.20	
	2002 – 2003		5.22	

* Log Mean Squared Difference between smoothed value and raw value in reference year.

** Ratio of Log Mean Squared Differences between smoothed single SMR and separately smoothed SMR numerator and denominator.

For all three smoothers (space, time and space-time) the correlations between the separated smoothing approach and the single value smoothing method are close to 1. As expected, the biggest differences in performance occur in the stability of the spatial smoother, for which the estimator with separately smoothed numerator and denominator has a noticeably smaller LMSD score than its single SMR counterpart. However, the predictive power of the separated and single SMR spatial smoothers is virtually identical.

Table 4.13: Predictive power of space smoother where numerator and denominator are processed separately compared with space smoother derived from single SMR value.

Method	Timescale	LMSE*	LMSER**
Separate smoothing of numerator and denominator of SMR	1999–2000	6.20	1.001
	2000–2001	6.17	1.001
	2001–2002	6.21	1.002
	2002–2003	6.22	1.001
Best space smooth from a single SMR value	1999–2000	6.20	
	2000–2001	6.16	
	2001–2002	6.20	
	2002–2003	6.22	

* Log Mean Squared Error.

** Ratio of Log Mean Squared Errors between smoothed single SMR and separately smoothed SMR numerator and denominator.

Table 4.14: Stability of time smoother when numerator and denominator are processed separately compared with time smoother derived from single SMR value.

Method	Timescale	Correlation coefficient	LMSD	LMSDR
Separate smoothing of numerator and denominator of SMR	1999-2001	0.999	4.44	1
	2000-2002	0.999	4.37	1
	2001-2003	0.999		
Best time smooth from a single SMR value	1999-2001		4.44	
	2000-2002		4.37	
	2001-2003			

* Log Mean Squared Difference between smoothed value and raw value in reference year.

** Ratio of Log Mean Squared Differences between smoothed single SMR and separately smoothed SMR numerator and denominator.

Performance of separated and single-value SMR time smoothers is virtually identical. We would expect this since these estimators take no account of spatial interactions.

Table 4.15: Predictive power of time smoother when numerator and denominator are processed separately compared with time smoother derived from single SMR value.

Method	Timescale	LMSE	LMSER
Separate smoothing of numerator and denominator of SMR	1999-2001	6.16	1
	2000-2002	6.16	1
	2001-2003		
Best time smooth from a single SMR value	1999-2001	6.15	
	2000-2002	6.16	
	2001-2003		

* Log Mean Squared Error.

** Ratio of Log Mean Squared Errors between smoothed single SMR and separately smoothed SMR numerator and denominator.

Table 4.16: Stability of space-time smoother when numerator and denominator are processed separately compared with space-time smoother derived from single SMR value.

Method	Timescale	Correlation coefficient	LMSD	LMSDR
Separate smoothing of numerator and denominator of SMR	1999-2001	0.980	3.16	0.98
	2000-2002	0.982	3.00	0.98
	2001-2003	0.983		
Best space-time smooth from a single SMR value	1999-2001		3.23	
	2000-2002		3.08	
	2001-2003			

* Log Mean Squared Difference between smoothed value and raw value in reference year.

** Ratio of Log Mean Squared Differences between smoothed single SMR and separately smoothed SMR numerator and denominator.

Finally, we come to the hybrid space-time smoother. Once again, differences between the separated and single-value SMR smoothers are marginal in this case, and driven by the effects on the spatial component of the estimator.

Table 4.17: Predictive power of space-time smoother when numerator and denominator are processed separately compared with space-time smoother derived from single SMR value.

Method	Timescale	LMSE	LMSER
Separate smoothing of numerator and denominator of SMR	1999-2001	6.11	1.01
	2000-2002	6.13	1
	2001-2003		
Best space-time smooth from a single SMR value	1999-2001	6.08	
	2000-2002	6.11	
	2001-2003		

* Log Mean Squared Error.

** Ratio of Log Mean Squared Errors between smoothed single SMR and separately smoothed SMR numerator and denominator.

It is important to remember that the space smoothers described here are very simple, and are driven by basic binary neighbourhood contiguity. Results obtained by Kafadar (2004) suggest that a more sophisticated spatial smoother using distance weighting might reduce the effect of population disparities further, and other methods of spatial smoothing (for example adjusting for population size) may also ameliorate this effect. The improvement in population consistency that will arise from using Super Output Area geographies will also reduce this problem.

It is clear from the existing results that space-time smoothing brings the best benefits in terms of urban and rural areas of the simple methods tested so far. Although there is little difference in the predictive power of separated and single-value smoothers here, smoothing the numerator and denominator separately is probably the most robust approach to take if spatial smoothing is used on its own or as a component of a space-time smoother. Separation of numerator and denominator does appear to have a positive effect on the stability of the spatial smoother without compromising predictive power.

5: Conclusions and recommendations

In this final part of the report we concentrate on the key findings of our preliminary investigation and their implications. We summarise the performance of the smoothing solutions evaluated and then review the limitations of the approaches investigated to date and indicate some key directions for further work in this area. Finally, we present our initial recommendations to Social & Health Analysis and Reporting Division (SHARD) on the production of smooth SMRs by the Office for National Statistics.

Summary of performance

The best space-time smooth identified in our tests combines a three year time smooth with equal weights across the years with a spatial moving average using weights that differentiate between first and second order neighbours, with slightly higher weight for the first order neighbours relative to the second order ones. The best time smooth and the best space smooth are the individual time and space smoothing components of this composite estimator.

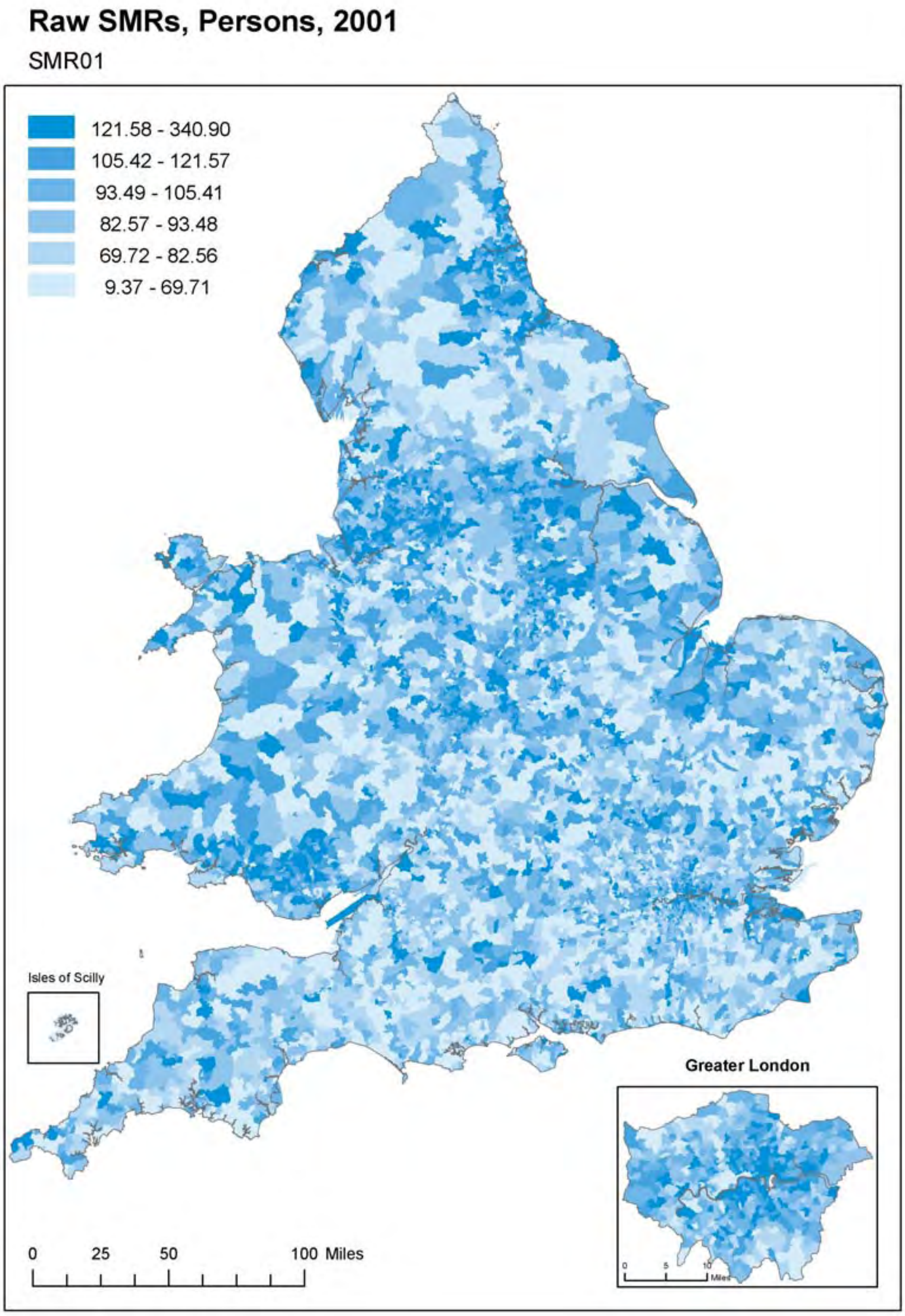
This estimator has three main strengths: it recognises the importance of accounting for spatial and temporal structure in estimating mortality ratios; it has conceptual simplicity, being is easy to explain and present and benefits from computational simplicity, as it is easy and quick to implement for large numbers of areas and time points.

Figures 5.1 and 5.2 illustrate the visual impact of smoothing in a pair of maps: one of raw values (5.1) and one of smooth values (5.2) produced with the best space-time approach identified, for the year 2001[10] . It is clear from a visual comparison of the two maps that, on a global scale, smoothing makes the geographic pattern of variation more visible and that the contrast between extreme values is more easily apparent.

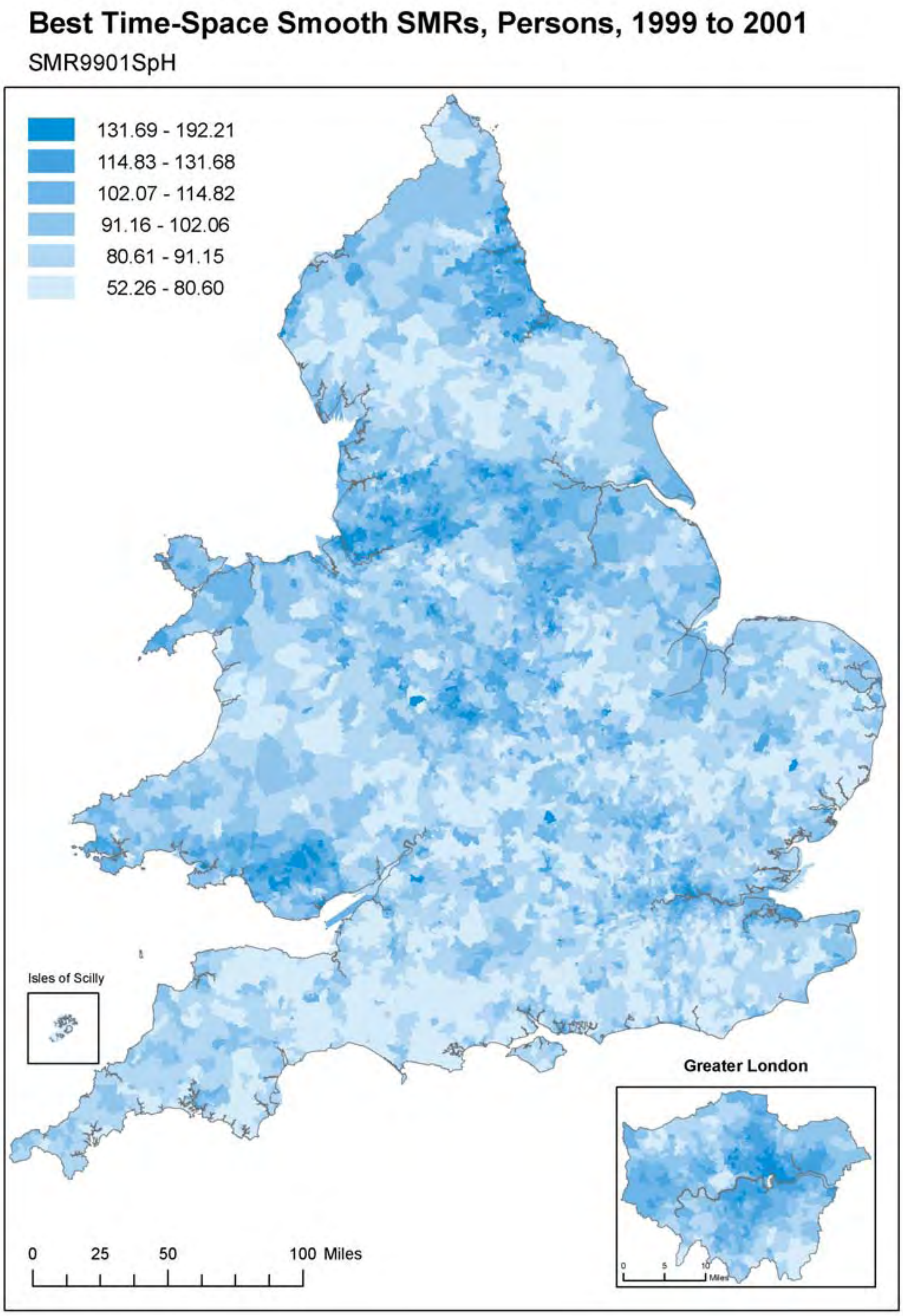
Appendix F contains summary information and maps for the raw and spatially smoothed five-year SMR, not included in the evaluation. As the five-year SMRs are a series of time smoothed data by construction, the alternative we show is that of spatially smoothing them with the best spatial smoothing solution identified above.

[10] The other two pairs of raw and smoothed data (2002 and 2003) are included in Appendix E.

Map 5.1: Raw data – 2001 Standardised Mortality Ratios for all persons.



Map 5.2: Best space-time smooth of Standardised Mortality Ratios from 1999 to 2001 for all persons.



The way forward

We have identified some basic methods of smoothing over time and space that appear to perform well against our evaluation criteria when applied to ward-level SMRs. We have implemented these methods using a suite of software packages, and produced some preliminary outputs using ONS mortality datasets.

The main strengths of the methods evaluated in this report relate to their conceptual and computational simplicity. The methods have, however, a number of limitations: they are built with arbitrarily chosen weights, do not produce any estimates of uncertainty for the smooth values obtained and do not account for variations in population across areas except in a very basic manner through the separate smoothing of the death count and population count and their combination after smoothing.

When applied directly to rates (rather than to the counts of deaths or of population at risk), the precision of the observed rates in each area remains unaccounted for. Also, linear smoothing methods such as these may be sub-optimal for data that do not follow a normal distribution.

We need to remain aware of the subjective nature of the choice of smoothing technology, as there is no universal optimal solution (Cowling et al., 1996), hence the need to explore and evaluate alternatives. This is of great consequence for the credibility of the estimates obtained and emphasises the importance of the need to convey the correct meaning of the 'smoothed' data to the user.

Other smoothing techniques

Alongside the linear smoothing approach used here, the spectrum of smoothing techniques includes a number of other possible solutions: kernel smoothing, non-linear smoothing, parametric and semi-parametric smoothing, empirical and hierarchical Bayesian smoothing, as discussed in chapter two in the "Data and techniques" section.

The limitations in the methods explored in this stage of the analysis could be addressed in a number of ways. To avoid the problem of varying populations across areas, weights based on population could be used (for instance by incorporating sufficient temporal or spatial lags to ensure that a minimum population of risk is used). The weights could also be specified to take the variance in each area into account (for example by making them proportional to expected counts). Choices of smoothing parameters such as those involved in kernel-based methods could also be beneficial.

Further work is needed in order to establish the suitability of such other smoothing techniques to the type of data analysed in this report and to individual level data.

Other methods of assessment

As indicated in chapter 3, the implementation of the cross validation framework would be made possible by the application of smoothing techniques to individual geo-referenced data. This would extend the current evaluation to cover both time and space.

Comparison with statistically 'principled' (i.e. model-based) approaches, in particular empirical and/or fully Bayesian methods would also be revealing, in particular through the quality measures that such alternative solutions can offer.

Alongside the clarification of the role and interpretation of the residuals in the theoretical framework proposed, an empirical exploration of the residuals (i.e. the differences between the raw and the smooth values) would be informative.

Recommendations

In general, we conclude that it is beneficial to apply some form of smoothing to annual SMR series in small areas if the objective is to extract the underlying variability in the data and to remove local noise. In fact, the method already used in the ONS for computing five-year SMRs at ward level belongs to the family of smoothing solutions that perform best across time. A small but unequivocal additional gain in both stability and predictive power can be obtained by smoothing over space the data produced by aggregating deaths over time.

Given that the further gain offered by spatial smoothing comes at the price of spatially varied performance and that further work will be needed to resolve this rigorously, we recommend that the time smoothing solution should continue to be applied as is for the present. However, it is worth noting that the move toward Super Output Areas as the preferred geography for reporting mortality statistics will reduce the impact of varying population size, and that the space-time solution is consistently the best performer in both urban and rural settings. It is important to note that this recommendation is based on the evaluation of the performance of similar smoothing solutions over three years rather than a direct evaluation of the existing five-year dataset and that, if more five-year SMR measures do become available, their direct evaluation would be required.

While most gain can be obtained from smoothing data over time and space, we also note that, compared to the raw data, well chosen spatial smoothing does bring a fair increase in stability and some increase in predictive power. It may, therefore, be worth considering space smoothing for noisy data when a time series is not available. The relatively lower level of gain is also augmented by the fact that pure spatial smoothing on annual data eliminates the need for revisions from one year to the next, which, although not directly relevant in this study, can be the case for some time smoothing solutions.

In general, we recommend that the numerator and denominator values of the SMR should be smoothed separately and then combined to produce a single measure, following the approach taken by others.

The findings remain relevant to the five-year SMRs only inasmuch as their three-year counterparts act as good proxies for the five year data. While we note that this is highly likely to be the case, further work on a time series of five-year SMRs would be necessary in order to confirm or reinforce their validity. Also, further exploratory work in the use of more sophisticated techniques and their implications, as well as consideration of alternative evaluation techniques would be beneficial if ONS wishes to develop best practice in the deployment of spatial and temporal smoothing methods for small areas.

Bibliography/reference

Banerjee, S., Carlin, B.P. and Gelfand, A.E., 2003, *Hierarchical modelling and analysis for spatial data*, Chapman and Hall / CRC.

Best, N. (2006). Personal communication.

Best, N.G., R.A. Arnold, A. Thomas, L.A. Waller, E.M. Conlon (1999). Bayesian models for spatially correlated disease and exposure data, in J.M. Bernardo, J.O. Berger, A.P. David and A.F.M. Smith (eds.), *Bayesian statistics 6*, Oxford University Press, 131-156.

Cowling, A; R. Chambers, R. Lindsay and B. Parameswaran (1996). Applications of spatial smoothing to survey data, *Survey Methodology*, 22.2:175-183.

Cressie, N.A.C. (1993). *Statistics for spatial data*, revised ed., John Wiley & Sons.

Elliott, P., M. Martuzzi and G. Shaddick (1995). Spatial statistical methods in environmental epidemiology: A critique, *Statistical Methods in Medical Research*, 4: 137-159.

Fotheringham, S., C. Brunsdon and M. Charlton (2000). *Quantitative geography: Perspectives on spatial data analysis*, Sage.

Gardner, M.J., Winter, P.D., Taylor, C.P. and Acheson, E.D., 1983, *Atlas of Cancer Mortality in England and Wales 1968 – 1978*, Chichester: John Wiley & Sons.

Getis, A. and J. Aldstadt, 2004. Constructing the spatial weights matrix using a local statistic, *Geographical Analysis*, 36.2, 90-104

Griffith, D.A. 1996. Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In Arlinghaus, S.L. (ed) *Practical Handbook of Spatial Statistics*, , CRC Press.

Grillenzoni, C. (2003). Non-parametric smoothing of spatio-temporal point processes, *Journal of Statistical Planning and Inference*, 128.1: 61-78.

Kafadar, K. (1994). Choosing among two-dimensional smoothers in practice, *Computational Statistics and Data Analysis*, 18: 419-439.

Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease, *Statistics in Medicine*, 15.23: 2539-2560.

Kafadar, K. (1999). Simultaneous smoothing and adjusting mortality rates in U.S. counties: Melanoma in white females and white males, *Statistics in Medicine*, 18: 3167-3188.

Kafadar, K. and P.S. Horn (2001). Smoothing, in El-Shaarawi, A.H. and W.W. Piegorsh (eds.), *Encyclopedia of environmentrics*, Wiley.

Knorr-Held, L. and G. Raβer (2000). Bayesian detection of clusters and discontinuities in disease maps, *Biometrics,* 56: 13-21.

MacNab, Y.C. and C.B. Dean (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates, *Biometrics* 57.3: 949-956.

Marsh, C (1988). *Exploring data: An introduction to data analysis for social scientists*, Polity Press.

Mills, T.C. (1990). *Time series techniques for economists*: Cambridge University Press.

Müller, H.G., U. Stadtmüller and F. Tabnak (1997). Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps, *Journal of the American Statistical Association*, 92.437: 61-71.

Simonoff, J.S. (1996). *Smoothing methods in statistics*, Springer.

Smith, D.P. (1992). *Formal demography*, London: Plenum Press.

Tukey, J.W. (1977). *Exploratory data analysis*: Addison-Wesley.

Veugelers, P.J. and S. Hornibrook (2002). Small area comparisons of health: Applications for policy makers and challenges for researchers, *Chronic Disease in Canada*: 23.3: 100-110.

Waller, L.A., B.P. Carlin and H. Xia (1997). Structuring correlation within hierarchical spatio-temporal models for disease rates, in T.G. Gregoire et al. (eds.), *Modelling longitudinal and spatially correlated data*, Springer Verlag, 309-319.

Waller, L.A. and C.A. Gotway (2004). *Applied spatial statistics for public health data*, Wiley-Interscience.

Appendix A: Standardised Mortality Ratio Calculations

The SMR in any ward is the ratio of the ward crude death rate (CDR) and the modified direct standardised death rate (DSDR*). The CDR is the ratio of actual deaths to population, by ward. The DSDR* is the sum of age-specific death rates in the reference population applied to the size of the local population, across age groups:

$$SMR_i = \frac{CDR_i}{DSDR_i *} = \frac{\frac{d}{n}}{\sum \left(\frac{D_x}{N_x}\right)\frac{n_x}{n}} = \frac{d}{\sum M_x n_x} = \frac{\text{observed deaths}}{\text{expected deaths}} \qquad \text{(A1)}$$

where:

Ward level	totals	d	Number of deaths
(i = 1,..m)		n	Number of persons
	specific age groups	d_x	
		n_x	
National level	totals	D	Number of deaths
(reference		N	Number of persons
population)		M	Death rate
	specific age groups	D_x	
		N_x	
		M_x	

Overall:

$$\begin{aligned} D_x &= D_x^{female} + D_x^{male} \\ N_x &= N_x^{female} + N_x^{male} \\ n_x &= n_x^{female} + n_x^{male} \end{aligned} \qquad \text{(A2)}$$

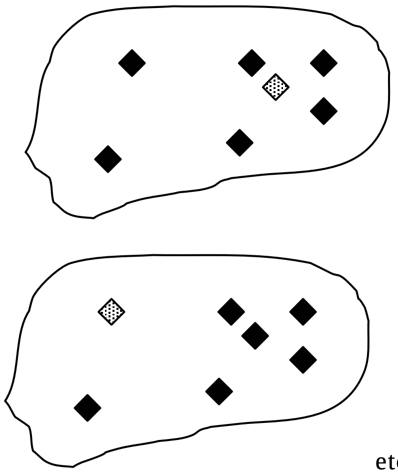
Appendix B: Spatial cross-validation framework

In this appendix we illustrate the key features of cross-validation that takes into account the spatial nature of the data, as a complement to the discussion on performance criteria in section 3.2.

Every area containing geo-referenced individual data can be divided (repeatedly) into a cross validation sample (cvs) and a sample used for the calculation of a smooth value (smooth). Spatial cross-validation is done by withholding one data point at a time and making a prediction at that point using data from the remaining points. Observed and predicted values and compared and errors are calculated. The method with the smallest error at most points is identified as the best technique.

Figure B.1 below illustrates an area containing a collection of geo-referenced points that are, one by one (dotted diamonds), withheld from the calculation of the smooth (based on black diamonds only).

Figure B1: Repeated spatial sub-setting of data



The sum of squared differences between the raw data in the cross-validation sample (cvs) and the smooth values calculated on the basis of the sample reserved for smoothing (smooth) can be decomposed in the following way:

$$\begin{aligned}
&\sum (y^{cvs} - \hat{u}^{smooth})^2 = \\
&\sum (u + e^{cvs} - \hat{u}^{smooth})^2 = \\
&\sum \left[\left(u - \hat{u}^{smooth} \right) + e^{cvs} \right]^2 \approx \\
&\underbrace{\sum \left(u - \hat{u}^{smooth} \right)^2}_{\text{Error of smoothed estimate}} + \underbrace{\sum \left(e^{cvs} \right)^2}_{\text{Noise term, which is independent of smoothing method}}
\end{aligned} \qquad \text{(B1)}$$

As the noise term is the same for any given *cvs* sample, comparing two smoothing techniques on the basis of $\sum (y^{cvs} - \hat{u}^{smooth})^2$ amounts to comparing the error of the smoothed estimates they produce, $\sum\left(u-\hat{u}^{smooth}\right)^2$. For any $\sum\left(e^{cvs}\right)^2$, the lower is the former, the lower is the latter. In other words, the closer is the smooth to the raw data, the better is the smoothing technique that produced it to retrieving the true underlying mortality, u.

Appendix C: Predictive power and over-smoothing

We expand here on the validity of predictive power as a safeguard against over-smoothing, in the context of the discussion of performance criteria in section 2.2.

At moment t, the estimated underlying mortality is, according to (5):

$$\hat{u}_{it} \equiv \underset{\substack{i^*,t^* \in S \\ t^*<t \text{ for } i^*=i}}{\hat{f}} \left(u_{i^*t^*}\right) + \hat{\delta}_{it} \tag{C1}$$

The raw data one step ahead can be written, by updating (1*), as:

$$y_{i,t+1} \equiv u_{i,t+1} + e_{i,t+1} \tag{C2}$$

and the corresponding underlying component can be written as an updated (2*):

$$u_{i,t+1} \equiv \underset{\substack{i^*,t^* \in S \\ t^*<(t+1) \text{ for } i^*=i}}{f} \left(u_{i^*,t^*+1}\right) + \delta_{i,t+1} \tag{C3}$$

Note that the function f in (C3) covers the past of $u_{i,t+1}$. Predicting $u_{i,t+1}$ is possible by using, *inter alia*, $\hat{u}_{it}$ and therefore $\hat{\delta}_{it}$.

If function f is miss-specified or if our estimate of the past as given by (C1) is too smooth and some of the information in the δ_{it} is lost, the correlation between $\hat{u}_{it}$ and $y_{i,t+1}$ (via $u_{i,t+1}$) is not as strong as it would have been, had over-smoothing not occurred.

Therefore, when comparing two smoothed series in terms of their predictive power, the technique producing the series with the higher predictive power is to be preferred, as it is the one that retains more of the underlying local information.

Appendix D: Time-space and space-time smoothing equivalence

Smoothing over both time and space entails smoothing over time data previously smoothed over space or smoothing over space data previously smoothed over time. In this appendix we show that, for the identified best performing smoothing options over time and over space, the two steps are equivalent.

The general forms for the two best performing smoothers are:

$$f(x) = \frac{1}{3}\left(x_t + x_{t-1} + x_{t-2}\right) \qquad \text{(D1)}$$

over time, equation (24), and

$$h(x_t) = 0.5x_{t,i} + 0.3\frac{\sum_j x_{t,j1}}{n_{j1}} + 0.2\frac{\sum_j x_{t,j2}}{n_{j2}} \qquad n_{j1} \neq 0, \quad n_{j2} \neq 0 \qquad \text{(D2)}$$

over space, equation (31).

Time smoothing of the spatially smoothed data amounts to calculating *f(h(x))*, while the spatial smoothing entails the calculation of *h(f(x))*, for each ward. We show that *f(h(x))= h(f(x))* and hence the two operations are equivalent:

$$f(h(x)) = \frac{1}{3}\left(h(x)_t + h(x)_{t-1} + h(x)_{t-2}\right) =$$

$$= \frac{1}{3}\left(h(x_t) + h(x_{t-1}) + h(x_{t-2})\right) = \qquad \text{(D3)}$$

$$= \frac{1}{3}\left[\left(0.5x_{t,i} + 0.3\frac{\sum_{j1} x_{t,j1}}{n_{j1}} + 0.2\frac{\sum_{j2} x_{t,j2}}{n_{j2}}\right) + \left(0.5x_{t-1,i} + 0.3\frac{\sum_{j1} x_{t-1,j1}}{n_{j1}} + 0.2\frac{\sum_{j2} x_{t-1,j2}}{n_{j2}}\right) + \right.$$

$$\left. + \left(0.5x_{t-2,i} + 0.3\frac{\sum_{j1} x_{t-2,j1}}{n_{j1}} + 0.2\frac{\sum_{j2} x_{t-2,j2}}{n_{j2}}\right)\right] =$$

$$= \frac{1}{3}\left[0.5\left(x_{t,i} + x_{t-1,i} + x_{t-2,i}\right) + 0.3\left(\frac{\sum_{j1} x_{t,j1}}{n_{j1}} + \frac{\sum_{j1} x_{t-1,j1}}{n_{j1}} + \frac{\sum_{j1} x_{t-2,j1}}{n_{j1}}\right) + \right.$$

$$\left. +0.2\left(\frac{\sum_{j2} x_{t,j2}}{n_{j2}} + \frac{\sum_{j2} x_{t-1,j2}}{n_{j2}} + \frac{\sum_{j2} x_{t-2,j2}}{n_{j2}}\right)\right] =$$

$$= \frac{1}{3}\left[0.5\left(x_{t,i} + x_{t-1,i} + x_{t-2,i} \right) + 0.3 \frac{\sum_{j1} \left(x_{t,j1} + x_{t-1,j1} + x_{t-2,j1} \right)}{n_{j1}} + \right.$$

$$\left. +0.2 \frac{\sum_{j2} \left(x_{t,j2} + x_{t-1,j2} + x_{t-2,j2} \right)}{n_{j2}} \right] =$$

$$= 0.5 \left[\frac{1}{3}\left(x_{t,i} + x_{t-1,i} + x_{t-2,i} \right) \right] + 0.3 \frac{\sum_{j1} \left[\frac{1}{3}\left(x_{t,j1} + x_{t-1,j1} + x_{t-2,j1} \right) \right]}{n_{j1}} +$$

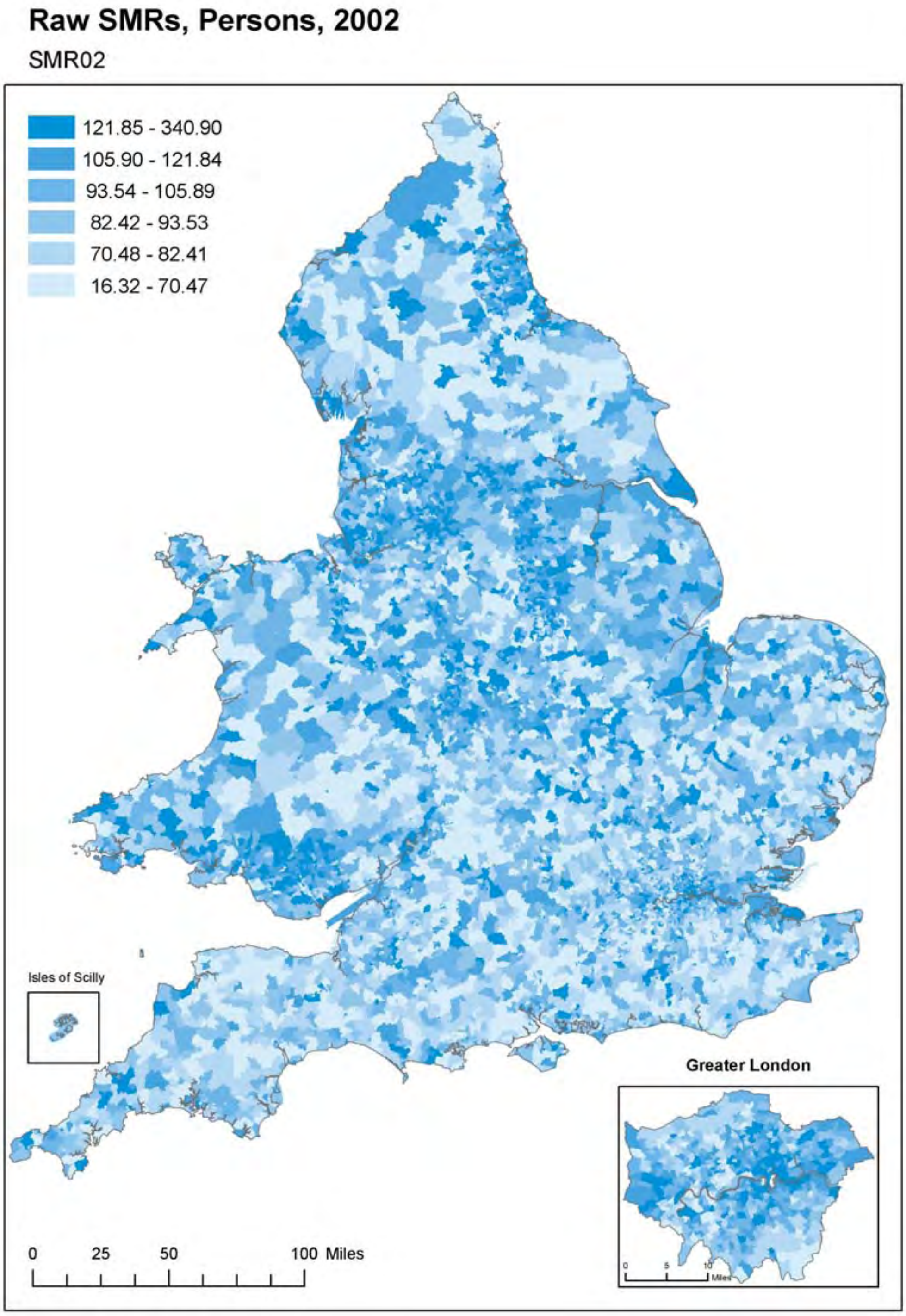
$$+0.2 \frac{\sum_{j2} \left[\frac{1}{3}\left(x_{t,j2} + x_{t-1,j2} + x_{t-2,j2} \right) \right]}{n_{j2}} =$$

$$= 0.5 f(x)_i + 0.3 \frac{\sum_{j1} f(x)_{j1}}{n_{j1}} + 0.2 \frac{\sum_{j2} f(x)_{j2}}{n_{j2}} = h(f(x))$$

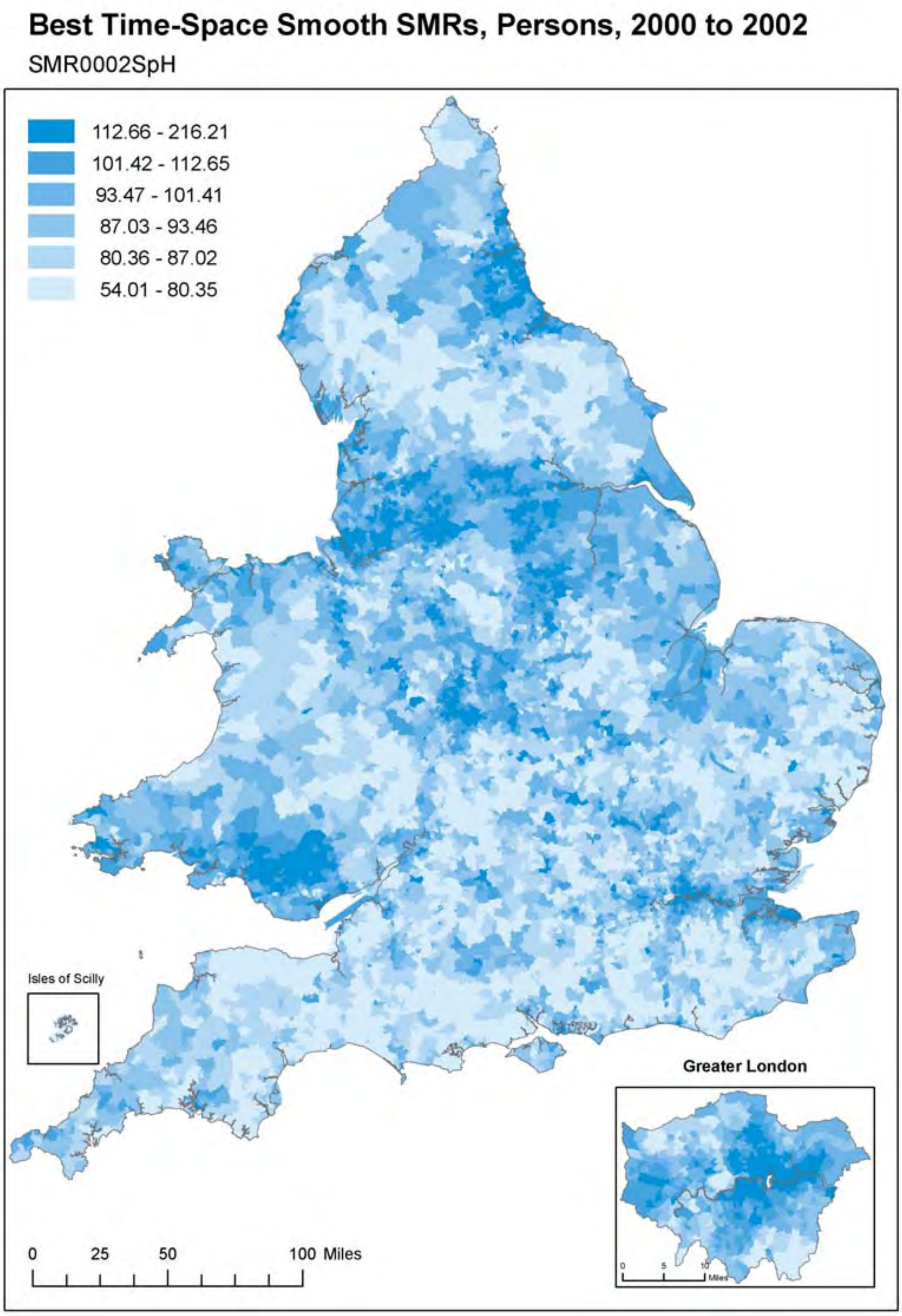
Appendix E: Maps of raw and smoothed data

The maps in this appendix depict the annual raw SMR data for 2002 and 2003 and their corresponding best space-time smooth values (2000-2002 and 2001 – 2003).

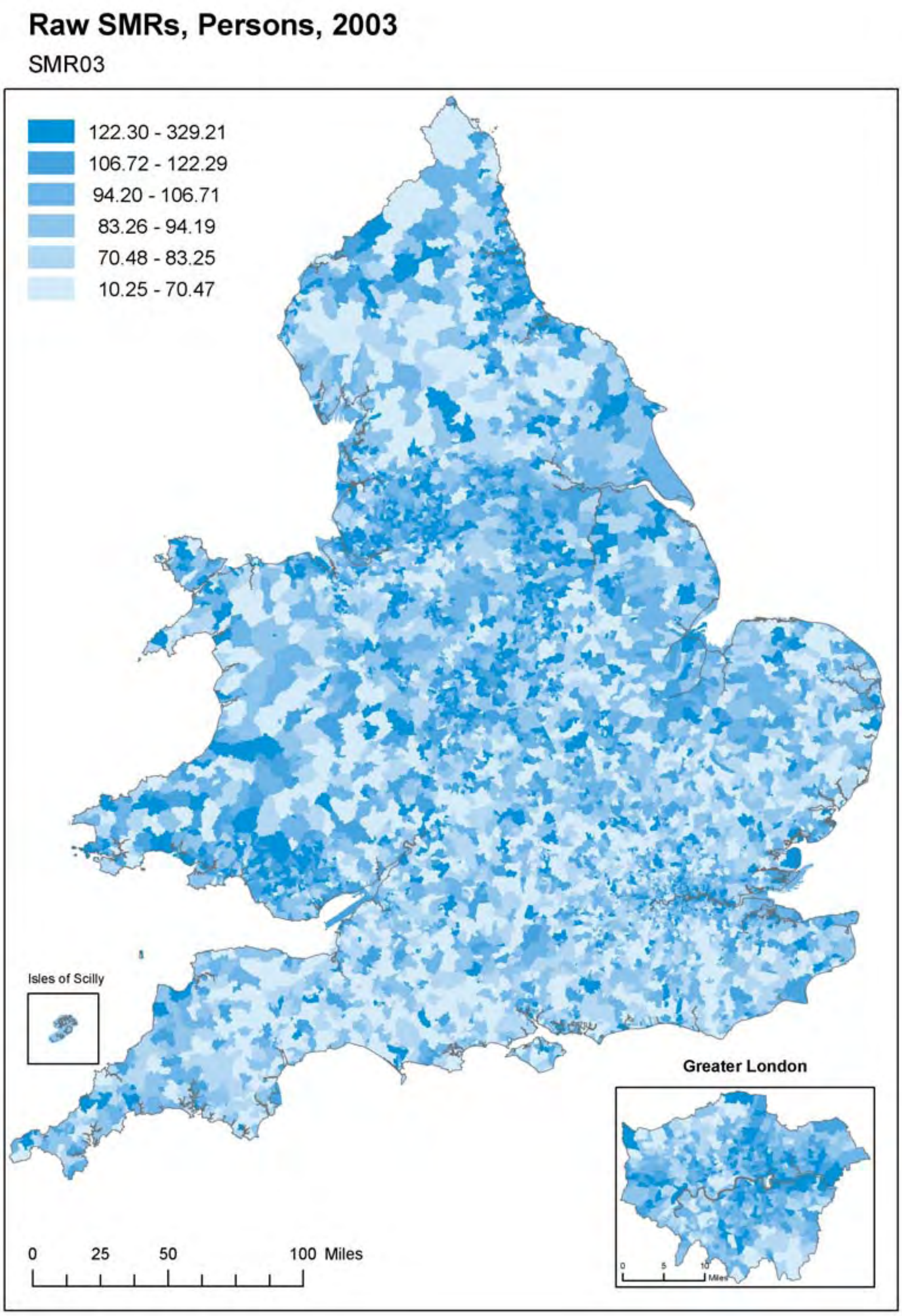
Map E1: Raw annual SMR data for 2002 - all persons in 2003 Statistical Wards.



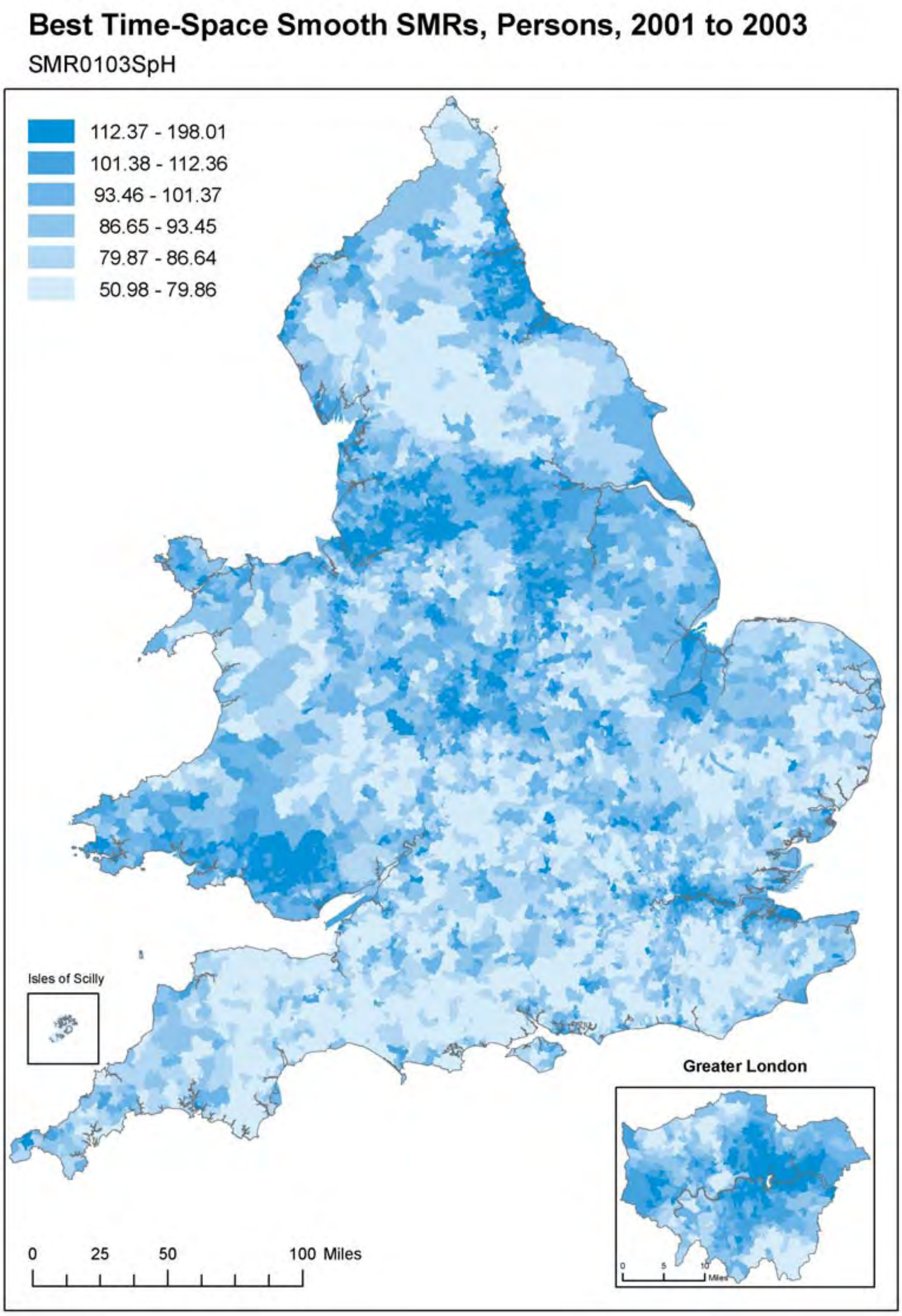
Map E2: Space-time smooth of SMRs for 2000-2002 - all persons in 2003 Statistical Wards.



Map E3: Raw annual SMR data for 2003 for all persons in 2003 Statistical Wards.



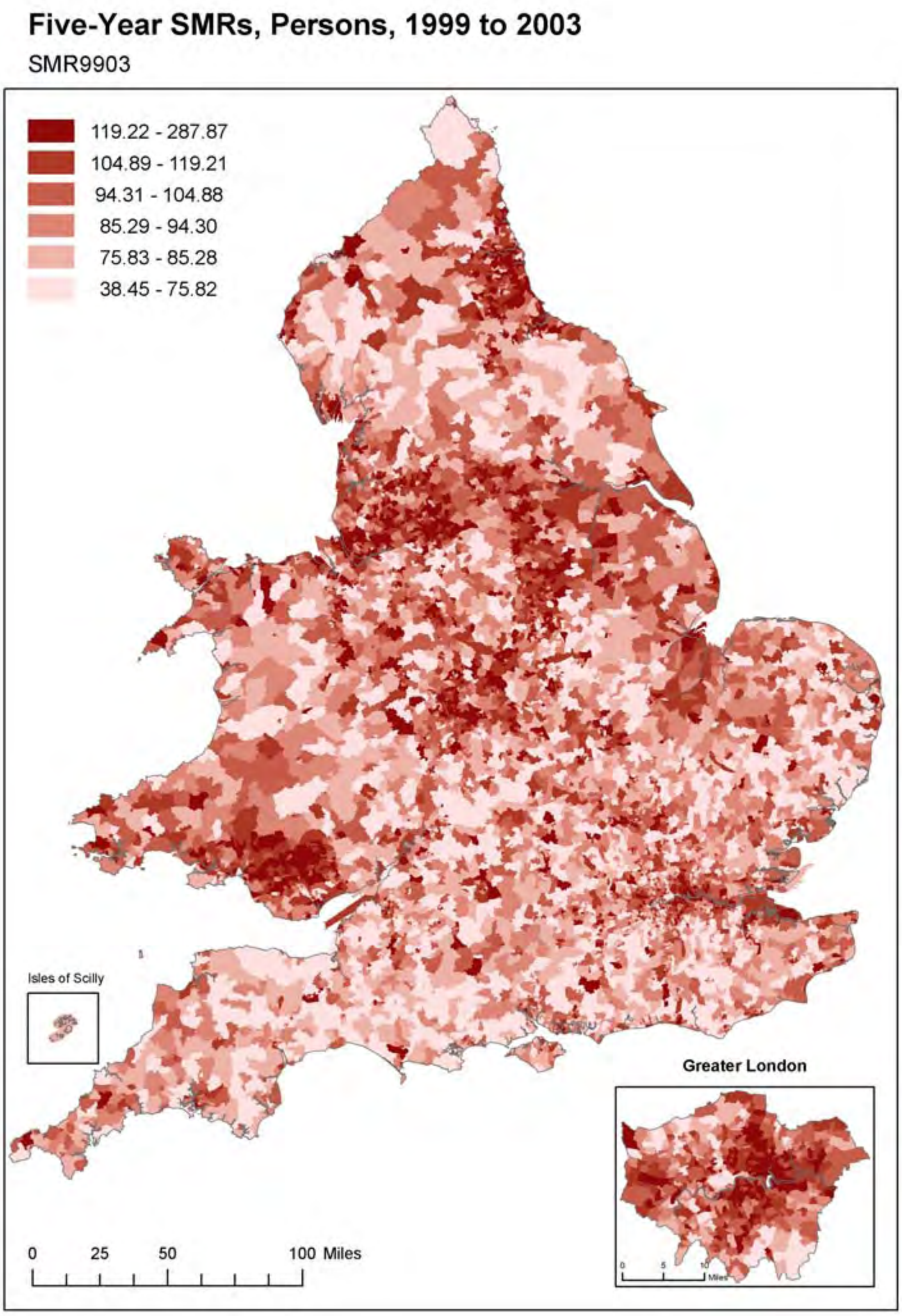
Map E4: Space-time smooth of SMRs for 2001-2003 - all persons in 2003 Statistical Wards.



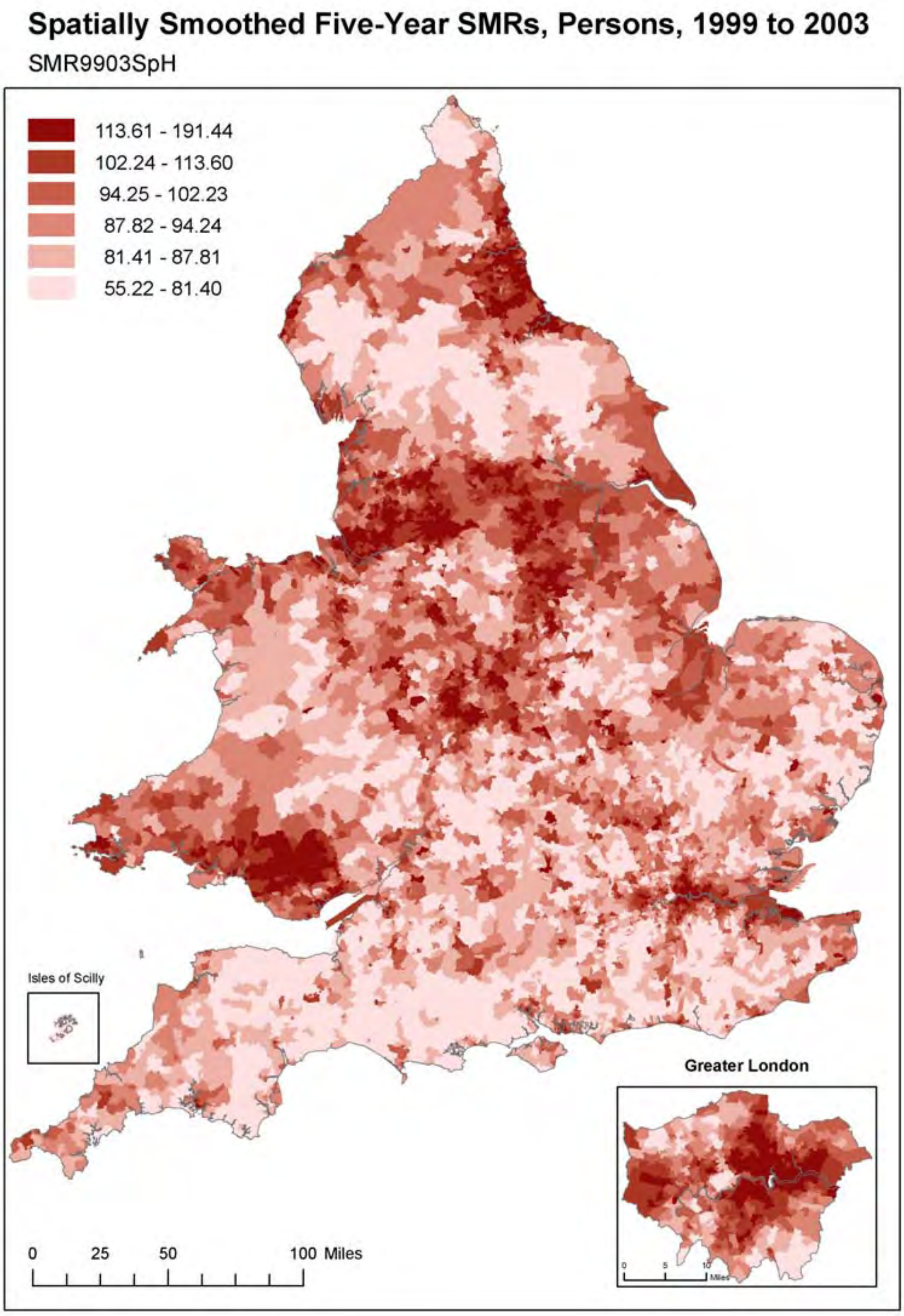
Appendix F: Five-year SMRs (1999 to 2003)

The maps in this appendix depict the five-year SMR data for 1999 to 2003, and the corresponding best space-time smooth, obtained by applying the best spatial smoothing method to data that have already been time-smoothed.

Map F1: Five year SMR data for 1999 to 2003 for all persons in 2003 Statistical Wards.



Map F2: Space-time smooth of SMRs for 1999-2003 - all persons in 2003 Statistical Wards.



Appendix G: Variable names

This appendix contains a list of names used for the variables included in the analysis. They correspond to the names used on the maps included in appendices K and L.

Raw data, annual:

SMR99

SMR00

SMR01

SMR02

SMR03

Time moving average, equal weights, 3 years (status quo):

SMR9901 (period 1999 to 2001)

SMR0002 (period 2000 to 2002)

SMR0103 (period 2001 to 2003)

Time moving average, declining weights (high memory), 3 years:

SMR3H9901 (period 1999 to 2001)

SMR3H0002 (period 2000 to 2002)

SMR3H0103 (period 2001 to 2003)

Time moving average, declining weights (low memory), 3 years:

SMR3L9901 (period 1999 to 2001)

SMR3L0002 (period 2000 to 2002)

SMR3L0103 (period 2001 to 2003)

Spatial moving average, one lag, equal weights (annual):

SMR99SE_1

SMR00SE_1

SMR01SE_1

SMR02SE_1

SMR03SE_1

Spatial moving average, one lag, high link (annual):

SMR99SH_1

SMR00SH_1

SMR01SH_1

SMR02SH_1

SMR03SH_1

Spatial moving average, one lag, medium link (annual):

SMR99SM_1

SMR00SM_1

SMR01SM_1

SMR02SM_1

SMR03SM_1

Spatial moving average, one lag, low link (annual):

SMR99SL_1

SMR00SL_1

SMR01SL_1

SMR02SL_1

SMR03SL_1

Spatial moving average, one lag, equal weights, enlarged (annual):

SMR99SE_12

SMR00SE_12

SMR01SE_12

SMR02SE_12

SMR03SE_12

Spatial moving average, one lag, high link, enlarged (annual):

SMR99SH_12

SMR00SH_12

SMR01SH_12

SMR02SH_12

SMR03SH_12

Spatial moving average, one lag, medium link, enlarged (annual):

SMR99SM_12

SMR00SM_12

SMR01SM_12

SMR02SM_12

SMR03SM_12

Spatial moving average, one lag, low link, enlarged (annual):

SMR99SL_12

SMR00SL_12

SMR01SL_12

SMR02SL_12

SMR03SL_12

Spatial moving average, two lags, high link (annual):

SMR99SpH

SMR00SpH

SMR01SpH

SMR02SpH

SMR03SpH

Spatial moving average, two lags, high link (annual):

SMR99SpL

SMR00SpL

SMR01SpL

SMR02SpL

SMR03SpL

Spatial moving average, two lags, high link, applied to time moving average, equal weights, 3 years:

SMR9901SpH

SMR0002SpH

SMR0103SpH

Time moving average, equal weights, 5 years (status quo):

SMR9903

Spatial moving average, two lags, high link, applied to time moving average, equal weights, 5 years:

SMR9903SpH

Appendix H: Time smoothing MATLAB function

This appendix contains the ***MATLAB*** function file that was used to calculate values smoothed across time. Comments are preceded by the symbol %.

```
function TS = tsmooth(Tj,w)

% Calculate a TIME SMOOTH


% { TS returns the time-smoothed value of a set of vectors that form matrix

% T and a vector of weights; the general form of Tj is Tj=[t1, t2,...tc],

% while w is a column vector of weights, with c elements. Thus, T(nxc) and

% w(cx1).


TS = (1./sum(w)).*(Tj*w);
```

Appendix I: Spatial smoothing MATLAB function

This appendix contains the MATLAB function file that was used to calculate values smoothed across space. Comments are preceded by the symbol %.

One spatial lag

```
function SpS = ssmooth_i(Sj,C,w)
% Calculates the SPATIALLY SMOOTHED value as WEIGHTED AVERAGE
%{ SpS returns the space-smoothed value of a vector or set of vectors (Sj),
% using a matrix of spatial interactions (C) and a vector of smoothing
% weights (w).

%{ The weighted average of neighbouring values is calculated using a
% row-standardised matrix of spatial interactions. The weights used here are
% the elements of the row-standardised matrix (can be a function of
% distance, e.g.). When C is a binary contiguity matrix, "xn" is just the
% simple average of neighbouring values.

Snj = [1./[sum(C')]'].*(C*Sj)

% The smooth values of Sj

SpS =(1./sum(w)).*([Sj Snj]*w)

% For any islands, where SpS would be not a number (NaN), due to the fact
% that sum(C) would be zero for them, the unobtainable smooth value (NaN) is
% replaced by the corresponding raw value:

n=length(Sj) %gives the upper limit for the index j below.
for j=1:n(1,1)
    if isnan(SpS(j,:))
      SpS(j,:)=Sj(j,:)
```

```
    end
end
```

Two spatial lags

```
function SpS = ssmooth2(Sj,C1,C2,w)
% Calculates the SPATIALLY SMOOTHED value as WEIGHTED AVERAGE
%{ SpS returns the space-smoothed value of a vector or set of vectors (Sj),
% using TWO matrices of spatial interactions (C1 for first order
% neighbours; C2 for second order neighbours) and a vector of smoothing
% weights (w).

%{ The weighted average of neighbouring values is calculated using two
% row-standardised matrices of spatial interactions. The weights used here are
% the elements of the row-standardised matrix (can be a function of
% distance, e.g.).

% first spatial lag
Snj1 = [1./[sum(C1')]'].*(C1*Sj)
  for i=1:5
    if isnan(Snj1(i,:))
      Snj1(i,:)=0
    end
  end

% second spatial lag
Snj2 = [1./[sum(C2')]'].*(C2*Sj)

  for i=1:5
    if isnan(Snj2(i,:))
      Snj2(i,:)=0
    end
  end
```

```
% The smooth values of Sj

SpS =(1./sum(w)).*([Sj Snj1 Snj2]*w)

% For islands, sumC1=0, so the result will be NaN; replace this by the
% corresponding raw value:

n=size(Sj)

for i=1:n(1,1)
    if isnan(SpS(i,:))
      SpS(i,:)=Sj(i,:)
    end
end
```