

EVALUATION OF STATISTICAL TECHNIQUES IN THE SCOTTISH INDEX OF MULTIPLE DEPRIVATION

Executive Summary

October 14, 2005

**Dr Alex McConnachie
Dr Chris Weir**

**Robertson Centre for Biostatistics
University of Glasgow**

This work was funded by the Scottish Executive, contract ref. PCM/04/13

The authors would like to thank the members of the Expert Advisory Group for their assistance throughout this project and the writing of the Final Report:

Prof Matt Sutton, Health Economics Research Unit, University of Aberdeen;
Dr Chris Dibben, Department of Geography and Geosciences, University of St Andrews;
Prof Hugh Gravelle, NPCRDC, Centre for Health Economics, University of York;
Dr Alistair Leyland, MRC Social and Public Health Sciences Unit, Glasgow.

Also, for their help during the project, the following members of the Office of the Chief Statistician at the Scottish Executive:

Ailie Clarkson
Robert Williams
Tracey Stead

1. INTRODUCTION

The Scottish Index of Multiple Deprivation (SIMD) 2004¹ is an area-based measure of deprivation defined for each of the 6,505 Scottish Neighbourhood Statistics (SNS) Data Zones². It is a combination of scores measuring 6 different aspects, or domains of deprivation:

- Current Income;
- Employment;
- Health;
- Education, Skills and Training;
- Geographic Access and Telecommunications;
- Housing.

Each domain score is constructed from a number of indicator variables, measured for each data zone throughout Scotland.

The SIMD 2004 is the result of a number of years' research into the measurement of deprivation at small area level, and bears close resemblance to other deprivation indices being used throughout the United Kingdom. The statistical methods upon which these deprivation measures are based have not previously been subjected to empirical assessment.

In November 2004, following the publication of the SIMD 2004, the Scottish Executive (SE) issued invitations to tender for the provision of an "Evaluation of the Statistical Techniques in the Scottish Index of Multiple Deprivation", in line with the long-term strategy for the measurement of deprivation in Scotland³. Particular areas under scrutiny were the use of Shrinkage and of Factor Analysis (FA), the methods of Exponential transformation and the weights used for the combination of domain scores into the final SIMD, and the estimation of uncertainty in the final SIMD scores and ranks.

A report of our findings was submitted to the Office of the Chief Statistician (OCS) at the SE in August 2005⁴; this Executive Summary gives an overview of the main findings and recommendations in that report.

We considered a number of modifications to the original algorithm for calculating the SIMD 2004, and compared the results obtained with those from the original method. One of the main methods for the evaluation of alternative methods was through the application of each method to 1,000 simulated indicator datasets (see Section 4.1 of the main report). Though not providing a complete solution to the issue of estimating uncertainty, the variations in each data zone's ultimate SIMD ranking over these simulated datasets were used to assess the relative degrees of rank uncertainty using alternative methods.

Throughout this project, the authors consulted with members of an Expert Advisory Group (EAG, see page ii), both in general and for some particular analyses; details of particular contributions are listed in Section 11 of the main report. The interpretation of the findings of this evaluation, and the recommendations made, are the views of the authors, and do not necessarily reflect the opinions of the EAG. On submission of this report to the funding

¹ Scottish Index of Multiple Deprivation 2004: Summary Technical Report. Scottish Executive, Edinburgh, 2004.

² Scottish Neighbourhood Statistics Data Zones: Background Information. Scottish Executive, Edinburgh, 2004.

³ Measuring deprivation in Scotland: developing a long-term strategy. Scottish Executive Central Statistics Unit, 2003

⁴ McCannachie A, Weir C. Evaluation of Statistical Techniques in the Scottish Index of Multiple Deprivation: Final Report.

organisation, the members of the EAG were given the opportunity to submit written comments on the final draft; these are included in Appendix F of the main report.

2. RECOMMENDATIONS

The purpose of this project was to evaluate the statistical methods used in the construction of the SIMD 2004, informally described as a “Health Check”. Other than a minor error in the original program code, the SIMD methodology could be said to have passed this checkup. We found little evidence that any of the methods used is invalid for the purpose of creating the SIMD or its constituent domain scores. We did, however, feel that there were some ways in which the statistical methods could be simplified or improved, and we have recommended a number of modifications reflecting this opinion.

2.1. SHRINKAGE

Shrinkage is currently applied in the calculation of two domains (Health; Education, Skills & Training) to guard against extreme data in small areas, by modifying indicator values towards the Local Authority (LA) average. In general, indicator variables tend to be shrunk by the greatest amounts in smaller data zones, since it is in these data zones that the level of precision for each indicator is least.

However, shrinkage involves a trade-off between variance and bias. One concern of using shrinkage is that indicator estimates are most biased in small data zones; if a small, but deprived data zone is present within an otherwise less deprived LA, this bias could result in the data zone being incorrectly ranked by deprivation⁵. We found some evidence of this “overshrinkage” of isolated pockets of deprivation; the use of a single (national) higher-level unit, or not using shrinkage at all, resulted in a slight redistribution of the most deprived data zones away from urban areas, where the majority of highly deprived data zones are concentrated.

The need to use shrinkage to guard against extreme values in small data zones may be negated by the use of large numbers of indicator variables, and the subsequent use of FA may have similar effects, so that the current algorithm effectively shrinks the data twice. The reduction in variability through the use of shrinkage was found to be small; not using shrinkage produced SIMD rankings in each data zone with less than 5% additional variation.

We therefore recommend that the shrinkage step of the algorithm be removed. It has little effect on the resultant indices and by shrinking towards LA averages, introduces a small bias that penalises data zones within otherwise less deprived areas. The application of shrinkage within some domains but not others does not constitute a consistent approach, and the use of Factor Analysis results implicitly in a degree of shrinkage.

2.2. FACTOR ANALYSIS

FA is currently used to combine indicators in three domains (Health; Education, Skills & Training; Geographic Access & Telecommunications) for which it is not possible to define the domain score as a simple sum of indicator variables. A one-factor model is used, which assumes the existence of a single latent variable, to which the indicator variables are linearly related in expectation.

⁵ The same effect could equally be seen in a data zone with low levels of deprivation within an otherwise deprived LA. However, the focus of the SIMD 2004 is more on identifying areas of high, rather than low deprivation, so this effect would not result in a miscounting of the number of highly deprived data zones within a LA.

The current method first ranks data zones with respect to each indicator variable, and then transforms these ranks to standard Normal variables. Each variable is standardised to the same distribution, but by ranking, the degree of separation on each variable between data zones is lost. A method of Generalised FA was explored, which retains the conceptual benefit of FA by assuming a single latent factor to which all indicators in a domain are linearly related, but the distribution of each indicator is modeled directly. When applied to the domains that currently undergo FA, we found only minor differences in the resulting SIMD ranks compared with the original method.

We recommend that a Generalised FA method be adopted. Though more difficult to implement than the current method, it removes the need to rank and transform indicator variables first, recognises the natural distribution of each variable and preserves the degree of separation between data zones with respect to each indicator variable.

We also recommend that Generalised FA be considered for the Current Income, Employment and Housing domains, as well as the three domains that currently undergo FA. This would have a minimal effect on the ranking of data zones with respect to these domains, but would result in a consistent methodology being applied to all domains. All domains could then be expressed as standard Normal variables.

2.3. EXPONENTIAL TRANSFORMATION

The methods used to combine domain scores into the final SIMD are designed to create an index of multiple deprivation that avoids “canceling out” should a data zone demonstrate opposing levels of deprivation on different domains. However, if the methods of Generalised FA were to be used for all domains, an alternative transformation would be possible that mimics the original method but does not involve ranking and therefore recognises the distances between areas on the original domain scores.

We recommend that domain scores, expressed as standard Normal deviates, are transformed by the function $f(x) = -\log(1 - \Phi(x))^6$, or similar, prior to their weighted combination to form the SIMD. This would fulfill a similar purpose to the current method in terms of avoiding “canceling out” of opposing levels of deprivation on different domains.

2.4. WEIGHTING

Changing the weights used to combine domains to form the SIMD has little effect. Whilst the weights currently used are therefore adequate, greater transparency could be achieved by explicitly separating the processes by which the weights are chosen, to reflect the prevalence and severity of each aspect of deprivation.

We recommend that the methods by which domain weights are derived is made more explicit, reflecting the importance of each domain in terms of its importance to those living in Scotland. Further research will be required into methods by which this could be achieved.

2.5. UNCERTAINTY

The most direct method to achieve rank uncertainty estimates would be to incorporate the entire SIMD 2004 algorithm within a Markov Chain Monte Carlo (MCMC) estimation procedure⁷ (e.g. using the software package WinBUGS⁸). However, the current methodology

⁶ The standard Normal cumulative distribution function, Φ , converts x (defined to have a standard Normal distribution) to a Uniform distribution, achieving the same result as ranking, but without removing the relative degrees of separation between data zones.

⁷ Goldstein H, Spiegelhalter DJ. League Tables and Their Limitations: Statistical Issues in Comparison of Institutional Performance. *JRSS A* 1996; **159**: 385-443.

involves the shrinkage of a large number of indicator variables and the repeated use of ranking of data zones, and would therefore be prohibitively complex and time consuming to fit.

Nevertheless, if the preceding recommendations were to be adopted, we believe that the algorithm could be incorporated into this framework. The only recommendation that does not result in a simplification of the method is the use of Generalised FA, for which MCMC methods are a natural means of estimation. The uncertainty in the final SIMD ranks could then be extracted as a by-product of the model fitting procedure.

We recommend that MCMC methods be used to produce the SIMD and component domain scores and ranks, with associated estimates of uncertainty.

2.6. OTHER RECOMMENDATIONS

We recommend that the CMR, CIF and Adults without Qualifications⁹ indicators be replaced by standardised ratios of the observed numbers of events to the expected numbers in each data zone, given the national age-sex distribution of events. This is a more widely used summary of event rates, and was found to have limited impact on the resultant SIMD ranks compared with the current methodology. If adopted, this would allow these indicator variables to be modeled more easily within the proposed Generalised FA technique.

3. CONCLUDING REMARKS

We have found that a number of simplifications to the current methodology could be made with little impact on the resulting deprivation indices. If anything, the current methods are slightly biased in favour of large urban areas, and a simplified algorithm could be viewed as more equitable. If one of the objectives of the current indices is to adopt simple and transparent methods where possible, then these simplifications would assist in this aim.

However, a more general method of FA could be used that would recognise the natural distribution of each indicator and preserve the relative differences between data zones on each variable. Such a method would be more complicated to apply, but no more difficult to understand. Generalised FA could be applied to all six domains without prior shrinkage of indicator variables, and would yield a similar ranking of data zones on the resulting SIMD. Such an approach is appealing in that each domain is treated in the same way.

The area in which the current indices could be improved the most is in the calculation of measures of uncertainty in the ranking of each data zone. This would allow the presentation of deprivation indices and summaries at composite area levels with associated confidence intervals, recognising the nature of such indices as estimates rather than truths. This uncertainty could be incorporated into resource allocation algorithms that are currently based on thresholds of deprivation levels, so that data zones close to the threshold would contribute towards the allocation attributed to the higher-level area. The necessary methods could be applied if the current algorithm were to be simplified, and should incorporate the proposed factor analysis methods with little difficulty.

⁸ Spiegelhalter DJ, Thomas A, Best NG, Lunn D. *WinBUGS version 1.4 user manual*. MRC Biostatistics Unit, Cambridge, available from www.mrc-bsu.cam.ac.uk/bugs:2003.

⁹ Scottish Index of Multiple Deprivation 2004: Summary Technical Report. Scottish Executive, Edinburgh, 2004.