

Work Session on Geographical Information Systems  
(Ottawa, Canada, 5-7 October 1998)

Item 6 of the provisional agenda

**THE USE OF GIS FOR THE SPATIAL AND TEMPORAL INTEGRATION OF UK  
STATISTICS**

Submitted by the Office for National Statistics, United Kingdom <sup>1</sup>

INVITED PAPER

---

<sup>1</sup> Prepared by Alistair Calder.

## SUMMARY

1. The Office for National Statistics has responsibility for collecting and producing statistics for a wide range of demographic, social and economic topics across much of the UK.

Recently a growing desire to integrate statistics from across these subject areas and to some extent with those from outside agencies, combined with customer demands for more flexible and responsive outputs, have led to a fundamental review of how ONS data is referenced. An effective referencing strategy for ONS needs to enable the production of accurate, responsive, flexible outputs as well as coping with the problems posed by the UK's complex and ever changing geography.

2. The strategy that is currently being investigated is a move away from the traditional building block approach to a solution where individual survey observations are allocated accurate co-ordinate references. This type of referencing, combined with the use of GIS and digital boundaries, should provide ONS with the flexibility and responsiveness to change required. This paper describes the problems associated with the existing referencing system, the strategy which has been proposed to replace it and the benefits which should flow from this approach.

## I. INTRODUCTION

### I.1 The need to integrate statistical data in the UK

3. The organisation of responsibility for statistical production in the UK is relatively complex with a number of separate government departments, sometimes split on national boundaries, collecting, managing and producing statistics in their own topic areas. The creation of the Office for National Statistics (ONS) in 1996 pulled together statistical production from a number of previously separate departments but the picture still remains complicated.

4. Nonetheless, as a result of its wide remit ONS needs to deal with statistics across a wide range of topics and collected from a wide range of sources - from surveys, from registration of events, from the census and from external agencies. Data is collected for individuals, for households, for businesses and for a wide range of aggregated areas.

5. Some statistics, notably environmental, transport and agricultural statistics, remain outside the remit of ONS but there is a growing awareness of the value of, and public desire for, the increased comparability and integration of statistical outputs across government. Moves are being made to simplify public access to national statistics, notably with the establishment within ONS of Statbase - a integrated database and dissemination mechanism (via the Internet) for statistics from across government.

6. Such pulling together of statistics from across topics and the growing pressure to try and integrate data offers tremendous opportunities but has also highlighted some of the weaknesses of our current referencing systems.

7. It has become clear that these systems do not offer the flexibility which is going to be required over coming years, particularly with respect to the integration of data across geographies and the ability to respond to change.

8. The need to find a better solution has recently lead to a fundamental review of how ONS data is geographically referenced. The group set up to investigate the options is currently assessing a radically different approach to referencing based upon co-ordinate references.

9. Before discussing this strategy in more detail it is worth considering the key problems relating to geography which impact on statistics in the UK.

## **II. PROBLEMS IN INTEGRATING UK STATISTICAL DATA**

10. The production of statistics in the UK is hindered by two fundamental problems related to the geographic base :

### **The existence of many different geographies**

11. Administrative geographies in the UK are complex. A series of reorganisations of the UK administrative structure over recent years as well as a tendency to invent new geographies at every opportunity has resulted in a complex web of boundary sets. Some administrative geographies nest within each other or can be used as building bricks for other geographies but many do not. Critically the postal geography, commonly used as a referencing base, does not fit within any administrative geography. Dealing with the issues of complex mismatched boundaries when trying to compare between datasets is certain to remain a key issue for statistics in the UK.

### **A high level of boundary change**

12. The second major problem in the UK is that boundaries, at all levels, are prone to change. Changes to some geographies are frequent, unpredictable and complex. Not only do high level boundaries move but the smaller building blocks from which other areas are constituted are also often subject to change. Some higher level boundary sets are defined in terms of frozen building blocks, some move as their constituent parts are subject to local boundary change.

13. Both of these problems lead to a requirement to continually re-cast statistics between areas - either to allow comparison between geographies or to provide new statistics for areas which have been subject to boundary change.

## **II.1 Current approaches to integrating data**

### **Re-casting statistics using building blocks**

14. The most common approach to re-casting data between geographies and producing statistical outputs for a variety of geographies is to adopt a low level building block to which the source data is allocated. Observations of events or surveys are allocated to, and statistics produced at, this building block unit. To produce statistics for different geographies these small building blocks are then allocated to higher geographies on a best-fit basis.

15. When statistics have to be produced on a different set of areas to allow comparison between different subjects the building blocks can be re-combined in a different order to represent the new geography. As boundaries change, the building blocks can be shuffled from area to area - providing new statistics. This is an obvious and easy to implement solution and is essentially the approach that has been adopted to date for the maintenance of UK statistics. And this is a

reasonable approach - it is simple, easy to implement and (as long as there is a significant difference in the relative sizes of the base and output unit) it produces reasonable results.

### **Problems of the building block approach**

16. However this approach is not ideal - there are some problems which arise from the method itself and some which are peculiar to the geography of the UK.

17. Firstly, the process of best-fitting using a building block necessarily introduces a degree of error into the results. Sometimes this is acceptable and the degradation of quality resulting from the best-fit process is not significant. However, where the difference in relative size of the source and output unit is relatively small, the error can make the comparison suspect. This means that this technique is usually not suitable for producing statistics at low levels - the noise introduced by the best-fit process destroys the quality of the source statistics. Equally a best-fit solution makes it impossible to identify or respond to very small changes, either in the boundaries of in the data.

18. Secondly, the best-fit method depends upon the existence of a suitable and relatively stable base unit. It is in this respect that the referencing systems currently used within ONS are beginning to show their deficiencies.

19. The low level building block which has been most commonly used in the UK is the postcode - a 7 or 8 character code designed and used for the delivery of mail. Postcodes vary in size - they can represent large single addresses such as businesses - but a size of around 15 households is common. The postcode is small, well established and commonly known and used in the UK. Within ONS the vast majority of data is coded to postcode. Other geographies are built either directly from postcodes or from larger units - but almost always with the postcode as a linking mechanism.

20. However the origin of the postcode as a operational tool for the delivery of mail results in a number of problems.

### **Problems arising from the use of postcodes**

21. Firstly, postcodes are not designed as a base geography and are not stable - new postcodes appear, old ones are terminated, often to be reused again sometimes in a different place. Control over the postcode is entirely at the hands of the Royal Mail (our mail delivery agency) and their location is neither logically locatable nor formally defined. Critically no real unit postcode boundaries exist - unit centroids are the only real source of geometric data for postcodes.

22. Secondly, since postcodes are designed for the delivery of mail they do not cover the whole of the land area. Although postcodes are small in urban areas they can be much larger in rural areas and some areas are simply not covered by postcodes at all. Although for demographic, social or business statistics this is relatively unimportant for environmental or agricultural statistics is a fundamental problem.

23. Most critically though, postcodes do not map exactly onto the main geographies used within ONS. Postcodes do not nest within administrative boundaries so any link from postcode to higher geographies are necessarily on a best-fit basis.

24. With these fundamental problems in mind - the complexity of UK geography, the tendency to change and our current dependence upon a base unit which brings with it a whole set of problems - it is now possible to discuss the approach to referencing which is being considered as a replacement.

### III. THE PROPOSED NEW REFERENCING STRATEGY

#### Summary of the method

25. The fundamental idea behind the proposed strategy is to make use of co-ordinate references as the key method of referencing observed data. When an event occurs or a household or business is surveyed the location of this event will be recorded and allocated a grid reference (an X,Y co-ordinate in the British National Grid) - the mechanisms involved in this allocation are described in more detail below. This grid reference can be used in the allocation of the observation to an area (and thus the production of statistics) but can also be stored away for future use.

26. The advantage of this solution in GIS terms is relatively obvious. Using simple point-in-polygon techniques it is possible to recut the data with any digital boundary set. By building a library of digital boundaries and mechanisms for allocating points to these boundaries we will gain almost complete flexibility. The availability of data referenced at such a low level also gives us access to much more sophisticated spatial modelling techniques than have previously been available.

#### In more detail : Allocation of grid references

27. The key to the allocation of grid references to captured data is a single digital dataset - ADDRESSPOINT. ADDRESSPOINT is maintained by the Ordnance Survey (OS - our national mapping agency) and provides a grid reference for every address in the Great Britain. It is the availability of this dataset, combined with developments in GIS and computing power which makes the proposed method of referencing a possibility.

28. When an observation is made the address of the household or business is captured. This address can then be matched against an centrally maintained directory based upon ADDRESSPOINT and a high quality grid reference picked up. The mechanism through which addresses are matched to the central database will allow for a degree of part and fuzzy matching of address.

29. In some cases, however, either because of the cost of data capture or because of issues of confidentiality, it will not be possible to use address. In such cases (or where a match on address has not been successful) a match will be done on the postal code alone. A similar directory (Data-Point - again an OS product) provides a direct lookup between postcode and a centroid co-ordinate for the unit postcode. Again this directory will provide a grid reference which can be used for allocation to higher areas and stored away with the data.

30. The problems associated with postcodes as a base unit previously described do not really hold here - once an item of data has been allocated a grid reference future changes to the postcode system are of no importance. It is the grid reference, frozen in time, which provides the

geographic reference. The only disadvantage to using the postcode as our link to grid reference is that the postcode grid reference is necessarily of lower accuracy (in terms of the resolution of the unit and thus the dataset).

31. Given that all observations are allocated a grid reference, either from the address or the postcode and that this grid reference is stored with the data, it is now possible to consider the flexibility that this will provide, both in terms of outputs and of the added value that can be obtained from the data.

### **III.1 Advantages of the proposed grid reference based solution**

32. In truth most of the advantages here are self evident and based upon a single principle. The co-ordinate system in which grid references are captured is common to all themes and geographies. By applying different digital boundary sets and using simple point-in-polygon techniques to cut up the grid referenced data we can obtain virtually complete flexibility.

#### **The ability to produce accurate statistics**

33. Because the base unit is now a single co-ordinate (a point) the results produced should be accurate. Given a reliable source of grid references and digital boundaries it is possible to produce perfect results!

Note : It is important not to get too carried away when thinking about accuracy here - there will still be a degree of error included in any result produced. The grid reference is a point and so infinitely small but this itself is subject to error - in many respects it is akin to a building block at the resolution (in terms of preciseness and locational accuracy) of the grid reference. But let's not worry too much about that for now.

34. Although we will never really get perfect results it is possible to produce statistics at a much lower level and with more reliability than was previously possible. In addition, every cut you make on the data is based on the source data at its lowest level and is as accurate as the first - there is none of the degradation of quality introduced by the best-fit solution.

#### **Responding to change**

35. As boundaries change all that is required to produce new outputs is to recut the data points with a replacement boundary set. It should be possible to identify even small changes this way. This has potential to result in significant savings on what is often a time consuming and very costly process.

#### **Comparison across geographies**

36. It is obviously possible to recast the grid referenced data for any geography for which you have a digital boundary set. This will make comparison between topics much more straightforward as data can be easily transformed to a common base.

37. One good example of how this might help is shown by considering environmental statistics. Environmental statistics necessarily have a different geography from other types of data - often defined by landuse or natural or man made features not reflected in the administrative

geographies. To date, this has resulted in a clear break between topics. Demographic data (for example) is not produced at a low enough level to be of real value to those that deal with environmental issues - the best-fit process makes the figures unusable at lower levels. Similarly environmental statistics are often produced on geographies which are of no use to those that produce demographic statistics. The move to the use of co-ordinates to reference demographic data makes the link from observed demographic data to environmental factors much easier. There should, for example, be no problem in identifying households or businesses within a flood area, in a potentially polluted area or within an area subject to some kind of conservation control. Equally the relationship between medical data and environmental hazards could be much more easily analysed.

### **Comparison of change over time and the creation of time series.**

38. An equally attractive aspect of moving to the use of co-ordinates as a referencing system is that this solution is essentially time-proof. Once data has been captured and a grid reference attached its position is known forever. As boundaries change over time it is possible to provide accurate estimates of how old data would have appeared on current boundary sets. Equally estimates for current data can be made using old boundaries.

39. Critically it should be possible to identify subtle changes with much greater reliability than could be obtained using a best-fit system. Obviously the value of grid referencing data will become more and more apparent as a time progresses and an increasing amount of old data can be used for such analysis.

### **Stability**

40. One of the real benefits offered by this approach is the genuine stability offered by the method. Although it is always risky to make such statements it seems that there is a purity about linking data direct to the co-ordinate system which makes this type of referencing future-proof. The ability to compare directly with data obtained by remote sensing or to integrate co-ordinates obtained from Global Positioning Systems (GPS) are obviously interesting ideas now but it seems inconceivable that changes in technology or future requirements will make the adoption of such a simple method look like a bad decision.

### **ADDED VALUE BENEFITS**

41. The advantages described above relate to a greater flexibility in producing standard outputs and responding to change. Many of the advantages which are anticipated from this approach, however, are added value benefits, new ways of managing, analysing and presenting data which have not been available to us before. Although it seems likely that many of these benefits will only become evident as the use of grid referenced data becomes better established, there are already several obvious opportunities.

### **Production of non standard outputs**

42. The logical conclusion of working from grid referenced data is that it is possible to produce outputs on any base required. This might be a specific set of output areas required for analysis or by a customer but equally it could be a regular set of areas such as a grid. The use of

gridded data does not have the prominence in the UK that it has elsewhere but the ability to produce such outputs may obviously be of value.

### **Adhoc enquiries**

43. Because data is held at very low level it will be possible to make use the tools available within GIS systems to carry out adhoc enquiries on data. In its simplest form this would allow us to identify, for example, those households falling within a certain distance from a hazard or transportation route but much more complex enquiries could be carried out. Demand for this type of enquiry is, as yet, untested but this type of flexibility is certain to be of interest to many users of statistics.

### **Construction of spatial and statistical models and other statistical benefits**

44. Similarly the value of spatial modelling in analysing and extending the use of data is largely untested within ONS. The use of spatial modelling techniques for visualising, analysing and interpolating data is largely dependent upon access to low enough level datasets. Access to grid referenced data will now allow us to make proper use of modelling tools and these are currently being investigated further.

45. Access to lower level data and the ability to integrate across datasets is also expected to have major benefits in terms of statistical methodology. Of particular interest is the ability to compare and combine datasets at lower levels than has previously been possible and thus construct better quality sampling frames. It is anticipated that this will be of particular value in the creation of synthetic estimates and dealing with non compliance in surveys. Equally, integration across statistical topics will help in the development of systems of area classification. Although these applications are currently more difficult to tie down it seems certain that these will be amongst the most important benefits of the proposed strategy.

### **Advantages from our growing experience of address matching**

46. Although not strictly a part of the referencing strategy a service to assist ONS business areas in matching between addresses lists is likely to be a fundamental requirement of the system developed.

47. The lack of true administration registers in the UK has meant that address lists are not as commonly used in UK statistical processing as in some other countries. Nonetheless a number of potentially useful address lists do exist and our growing use of address as a link to grid reference and experience in matching between lists seem likely to make these an important new source of data. In the short term, address matching seems likely to be of benefit in building improved sampling frames, refining synthetic estimates and in matching data between various business registers held within government.

## **III.2 Problems associated with the proposed change**

48. The outline of our strategy presented here is necessarily simplified. A number of important issues and problems remain, not least whether the quality of the datasets available to us are suitable for this type of work. These issues are being investigated as part of a research programme

which will feed into a full business case. Most of the issues being considered are specific to our case but a couple of issues are of wider interest and are mentioned below.

### **Organisational issues of the implementation in business areas**

49. The proposed strategy represents a fundamental change to what are in many cases critical processing systems. There can be no room for error in introducing the new method of referencing in ONS business areas. An effort is being made, however, to make the process as close to a black box solution for business areas as possible. They will capture data (ideally moving to an address reference) and produce outputs in exactly the same way as they do presently. Systems for producing standard outputs should continue largely unchanged. The main difference will be that as part of their data coding process they will get access to a grid reference which can be stored away with their data for future use. It seems likely that any move to address level referencing will be introduced as new systems are developed rather than being part of a wholesale change to ONS systems.

50. As with any change to such systems gaining the commitment of individual business areas will be critical. It is hoped however that the potential benefits, and relatively low compliance costs will make an early move attractive to many areas. Critically in order that ONS gains maximum benefit from the change a central service will be established to promote the strategy and support business areas in all aspects of using GIS and grid referenced data.

### **Confidentiality**

51. The move to dealing with address level data and particularly the possibility of producing such flexible, potentially overlapping, outputs necessarily raises the issue of confidentiality. This is obviously an issue which needs to be dealt with - and controls will have to exist on what data is released. However, in many respects, there is nothing really new here - this is not an intrinsically geographic problem. Pressure to produce more flexible outputs necessarily require careful attention to issues of confidentiality from the point of view of data classifications as well as geography. Responsibility to ensure that outputs are not produced which compromise confidentiality rules currently rests with business areas - and will have to remain there.

## **IV. CONCLUSION - PROGRESS TO DATE AND PROSPECTS**

52. The programme of research considering the benefits and issues related to this proposal is currently underway. This programme is intended to feed into a full analysis of the systems required and potential benefits and a full business case at the end of 1998

53. Whether we adopt this strategy depends upon whether the benefits of accuracy, flexibility, responsiveness to change and the added value it offers outweigh the costs associated with purchasing OS data and the changes required to existing systems. We believe they do but our ability to quantify the benefits effectively remains to be seen. We already have some enthusiastic business users and others who are yet to be convinced. A promotional program to inform and seek the views of business areas is currently under way. We will know by the end of 1998 whether the strategy will be adopted for ONS. A number of other UK statistical departments are very interested in this initiative and are awaiting the result - if the strategy is adopted and proves to be successful it seems likely that a similar model will be considered for other UK departments.