



# **Developing Procedures for Articulating Policy-Related Evidence**

**Peter Bibby**

Department of Town and Regional Planning

University of Sheffield

**January 2005**

## **1 Introduction**

- 1.1 This note outlines work undertaken at the Department of Town and Regional Planning at the University of Sheffield (TRP) which attempts to consider how evidence might be articulated to guide assessment of policy, initially in relation to Defra's rural funding streams but ultimately with much wider application.
- 1.2 The starting point in developing evidence is that it must be germane to the needs of policy makers and to policy assessment. This involves 'translation' between the things that policy makers and what is typically thought of as data. Now this matter of translation might be thought of as involving a bundle of skills. Nevertheless, it might make sense to see whether there are embedded within this certain classes of generic tools which might assist in this translation. This gave rise to a series of possible operations and a series of experimental applications.
- 1.3 The type of approach focuses on adding value to existing data, typically by exploiting the logical relations between separate data sets. It does not assume a particular data substructure but is intended to exploit such data in such form as may be available.

## **2 The Foundations**

- 2.1 The foundation of the approach lies in the need to identify things that policy makers and policy analysts talk about with a focus on social or geographic objects. They may talk about 'areas with a dependence upon X'; they may talk about 'market towns' or 'villages'. These are not necessarily the areas for which data are collected. Crucially rather than attempting to assess the characteristics of a data rich statistical or administrative unit such as a ward, objects of discourse frequently refer to entities which do not correspond to areas for which data are collected.
- 2.2 Alternatively they may allude to objects whose nature may be contested (e.g. the 'Warwickshire housing market' or the 'southern crescent'-(an entity debated in preparing regional policy guidance for the North West). The requirement to identify such areas and to formulate policy with respect to them continually recurs. Typical questions might include
  - How is employment changing in the labour market area of Scarborough?

- Can we identify areas in the Pennines with a high dependence upon agricultural employment and a need for farm diversification?
- Can we identify distinct rural housing market areas across England? Is there evidence that land supply constraints is driving up house prices in such areas to a disproportionate degree?

2.3 Pertinent information may of course be available for areas poorly matched to the users' needs, or need to be estimated using quanta appearing in datasets for discordant geographic bases. This implies a need to manipulate data to make it apply from one set of areas to another set of areas (which is considered under the heading of areal interpolation). Frequently data may exist to form a very good approximation to a variable (e.g. the population of a particular small settlement), but the logical chain which links the natural language expression to potential estimators has not been specified. This is a core concern of the approach discussed here.

2.4 The approach introduced has an explicit philosophical underpinning whose practical implications are being pursued as a long-term project by TRP in the University of Sheffield. Ultimately, following Jubien (1993) it treats 'things' as bundles of infinitesimal elements of matter brought together by natural language (see Bibby 2005). For practical purposes this is approximated by methods which allow geographic objects to be constructed out of small tiles, analogous to the tesserae forming a mosaic pavement.

2.5 Tiles are constrained to take positions within a regular tessellation, forming a restrictive and limited subset of what might be identified by Jubien as 'things'. More specifically, for the purpose of the present project, areas of policy concern are constructed out of any combination of square tiles each representing an area on the surface of the earth of 100m by 100m (i.e. one hectare in extent). At least initially, a response to a question of substantive concern might be thought of as involving the specification of an area to be represented by a number of black tiles, against a white ground (egg where are areas within the moorland line? or where are settlements with a population of less than 500 persons?).

2.6 **Hectare** tiles are chosen simply for convenience. Conceptually tiles might be much smaller, though this would increase the computational burden. Hectare tiles seem appropriate for many purposes. Data for individual properties might be aggregated up to hectare cells; while data for units such as wards might reasonably be disaggregated to this level. It is important to appreciate that as part of the procedures developed, tiles are assigned characteristics, and that these characteristics depend upon objects that **impinge** upon them (which may be much smaller or much larger and which need not be fixed in space).

- 2.7 For present purposes, an object of concern is anything talked about - for example a person, a field, an organization, a belief or even a hectare tile. This definition includes but extends beyond all Jubiennesque 'things.' It need not necessarily be represented by a set of hectare tiles, but in whatever form may be convenient (e.g. a text string in a database, or a natural language expression held as a Prolog list). To say that objects may impinge upon tiles is not to say that they necessarily fall within or overlap them (though this may be the case), but simply to say that they have some form of connection.
- 2.8 This approach gives rise to a series of opportunities but inevitably introduces some new problems. First of all, it demands that there is some way of translating data from its 'native' form to a target representation of hectare tiles. This in turn requires ways of building rules to assign attributes to hectare tiles, (which incidentally makes use of methods of drawing understanding from natural language). Thirdly, it demands methods of drawing hectare tiles together to form new areal units and of generalizing the attributes of hectare tiles to higher levels. Fourthly it requires facilities to display and tabulate the characteristics of the areal units constructed. Finally, it seeks to integrate methods of drawing inferences from whatever information is available, defining and making use of relationships between data items.
- 2.9 These objectives can be achieved computationally in a variety of ways. Within the present project, a range of software has been used, but integration has been achieved using the logic-programming paradigm implemented in Prolog. In considering implementation of the approaches discussed here, the tile metaphor is central. It is important to realize that it stands apart from particular data structures (particularly those embodied within current Geographical Information Systems). It is chosen partly because it is ambiguous in two respects.
- 2.10 Firstly, while any tiles may be thought of as *mereological sums* irrespective of their spatial configuration (a circumstance not easily represented within proprietary Geographical Information Systems), tiles are constrained to occupy locations in a fixed lattice. Secondly, the tessellation metaphor might be seen from either a vector or a raster perspective. At the level of practical processing, it proves useful to shift between different representations, but the tesseral metaphor stands regardless of particular data structures.
- 2.11 The remainder of the body of this document says a little more about these principles, while specific points are taken further in a series of appendices.

### **3 Estimating Measures for Hectare Tiles: Aggregation and Disaggregation**

3.1 Conceptually, the first major step in the process outlined in para 2.5 involves assessing the characteristics of hectare tiles. In practice, this involves a range of procedures, but generically it will always involve three sub-tasks

- finding appropriate existing substantive data (regardless of spatial referencing)
- finding pertinent spatial data and recasting it in a form compatible with the hectare tessellation
- specifying the logical links between all relevant data and estimating relevant quantities at the tile level

3.2 Each of these subtasks obviously covers a diverse range of activities, but some examples are provided below.

#### **Finding appropriate existing substantive data**

3.3 The critical point here is that pertinent data need *not* be spatially referenced, but merely potentially linkable. Many data are postcoded at unit postcode level and so easily locatable. In some cases (such as CACI small area income estimates), the relation of the postcode to the objects of concern (households) is clear and their use straightforward. In other cases, such as the company information appearing in Financial Analysis Made Easy (FAME), the relation of a postcode to substantive data (such as company accounts) is much less obviously, and the extraction of relevant postcodes itself more problematic. In other cases, the relation of an object to a postcode is straightforward (when for example it is treated as an index of a unit of occupation), but the characteristics of that unit is locked up in a natural language expression such as 'The Marquis of Granby'.

3.4 Although the preceding paragraph illustrated the issue by reference to data that were directly or indirectly postcoded, the question of identifying pertinent substantive data is an issue of far broader generality. It is necessary to have regard to logical chains that might be used to add value to apparently placeless sample data (such as the 2000 Time Use Survey), to treatment of textual data as indexical signs (as above), to 'part-whole' relations (e.g. between shops and shopping centres, between historic farmsteads and hamlets).

### **Finding pertinent spatial data and recasting it in a form compatible with the hectare grid**

- 3.5 It is assumed that through the specification of some arbitrarily long or short logical chain it will be possible to associate some spatial reference (where appropriate) with the substantive data referred to above. The spatial data might take the form of point, line, polygon or raster data in a range of (proprietary) formats. Such data can usually be relatively easily converted into forms which may be used with products such as ArcGIS, Arcview or Idrisi.
- 3.6 For the purposes of display and many forms of geographic analysis, such systems provide ample facilities, though these prove either an insufficient basis for undertaking the type of flexible 'constructivist' analyses designed to identify the things which populate policy discourse (or prove impractical). This note will concentrate on the circumstances where further facilities are helpful, and logic programming proves useful.
- 3.7 Where data are available in raster form (e.g. Land Cover Map2000), or they are properly considered as polygons (e.g. the Countryside Character Area typology) and can be converted to a grid of hectare cells (of standard extent, origin and orientation), they may be converted to Prolog predicates characterizing attributes of hectare tiles. At present, this is achieved using a FORTRAN conversion program to generate facts of the form

tile (Q, K, M),

where Q indicates a unique code (considered further below) indicating the tile; K is a flag indicating the nature of 'the matter in the tile and M the mass or 'amount' of matter in the tile.

### **Specifying the logical links between all relevant data and estimating relevant quantities at the tile level**

- 3.8 Achieving the goal of estimating the attributes of a tile by reference to the objects and activities that impinge upon it, involves bringing together the pertinent substantive data and the spatially referenced data. Although the logical chains that tie the data together may be long, and the notion of impingement complex, it will be convenient at this stage to limit consideration to circumstances where an object for which there is substantive data falls within a tile or where a tile falls within such an object.
- 3.9 The former case, in which for example, a rain gauge or a guesthouse is located within a hectare tile aggregation is straightforward. Disaggregation of statistical measures relating to broad areas (such as Output Areas or wards) is obviously more contentious. The approach to interpolation

developed in TRP and applied in this study is considered in more detail in Appendix 1, but it may be appropriate here to comment on the methods ATT the conceptual level.

- 3.10 In the case of socio-economic data, interpolation may be regarded as being undertaken by reference to an underlying distribution of *facilities*. From this perspective, the world might be thought of as comprising facilities or more fully, *milieu-behaviour synomorphs* – or properties or parcels of land over which typical behaviours occur (for application to philosophy of social and geographic objects see Casati et. al. 1998).
- 3.11 The limits of the activity are the limits of the associated space (A football game would be a classic example, a church service, a car boot sale, a restaurant, and a dwelling house, while a village or a town centre provide more complex and debatable) examples. Policy discourse proceeds as if such units existed, though they may well be contentious (the extent of a foxhunt or a rave; 'dogging sites'). More generally as geographic scale increases, the applicability of the notion of synomorphy becomes less clear. (Many of the difficulties are evident in the idea of a place 'community').
- 3.12 The justification for interpolation is that the activity recorded in statistical aggregates is attributable to the activity associated with particular facilities. Postal addresses together with organization names serve as an index allowing the distribution of particular classes of facilities across tiles to be imputed and hence a finer scale geography of activity derived.
- 3.13 In the case of many facilities this imputation is straightforward. The organization name 'McDonalds Restaurants' within PAF gives a very strong indication of the precise nature of nature of the activity across a particular sort of space. This in turn implies that from sets of postal addresses we can infer the numbers and distribution of facilities, and using natural language processing, it is usually possible to infer the character of a facility.
- 3.14 The sorts of facility discussed above are typified by activities which include responding to mail and it is for this reason that they are indexed by Royal Mail's PAF. The idea of synomorphy can be extended beyond such examples- enclosures for stock may also be thought of as facilities, though these are not indexed by Royal Mail. Moreover, the notion of synomorphy itself becomes less clear at increasing scales. Thus, we are familiar with intensions that arise in the concurrent use of land in National Parks, for recreational, agricultural, and conservation purposes. Nevertheless, for extensive land-uses, the idea of synomorphy has some pertinence as particular land covers are associated with particular activities, and historic farm payments data might be used to explore this.

- 3.15 For the purpose of this project an approach has been developed to interpolating information about extensive land-using activities which involves the use of land cover information as a proxy for knowledge of the distribution of facilities. This has the advantage of providing comprehensive coverage without having to piece together diverse sources, but it lacks an authoritative index of facilities analogous to Royal Mail's PAF.
- 3.16 The most obvious source of data is Land Cover Map 2000, which is derived from satellite imagery. This in principle provides a measure of land cover for 25 metre tiles, though in practice there are many concerns about reliability. For this reason we have sought for the purposes of this project to explore the feasibility of integrating data first to draw better inferences about land cover (milieu) by taking account of context, and then develop methods of relating this to land-use (behaviour).
- 3.17 For this purpose it is been convenient to deploy a revised and simplified set of land cover categories
- Coastal & Estuarine
  - Inland Water
  - Bog and Heath
  - Woodland
  - Grassland
  - Cropland
  - Avegetal
- 3.18 Tiles are assigned to a class on the basis of LCM 2000, other cover sources (e.g. NIWT, OS Meridian mapping, land use information (e.g. individual Property Use, and contextual information (e.g. elevation, settlement morphology)). Having identified a set of land cover types they may be used to guide the interpolation of broader scale information about agricultural activity and agricultural output.(e.g. that obtainable from the June agricultural Census).

#### **4 Composing Derived Areal Units: Aggregation and Generalization**

- 4.1 The preceding paragraphs sketches out an approach to identifying all those tiles that meet selected criteria. As suggested previously, these criteria may be specified with greater or lesser degrees of complexity: -the range of possibility is discussed in Section 5 below. For the present, it is convenient to assume that pertinent tiles have been identified and to consider how they might be combined. Consistent with the overall approach, the selected tiles might be thought of initially just as a number

of parts- conceptually a mereological sum. If the parts are given unique identifiers, they may be represented in a list.

4.2 For our purposes, however, it is desirable to move beyond the list of parts and to consider their *form* i.e. to enquire whether any of the parts are contiguous and to build these parts into a number of discrete objects (such as distinct areas with a particular problem or which express particular opportunities). For example, let us say that there is a requirement to identify a rural settlement referred to as 'Howick' and to assess its population in 2001. This apparently trivial example is illuminating in that all the necessary data exist; but the possibility of answering it is hampered by inability to handle the natural language specification.

4.3 Consistent with the approach described above, the answer will take the form of an operation on a mereological sum of hectare tiles. Consider first how that mereological sum will be found. The natural language postal addresses in PAF indicate those which are associated with Howick. Each has a unit postcode, identified with a unique hectare tile, and so all properties addressed through L can be represented as a list. In practical application the list actually takes the form

[[54559551, 1, [howick]], [54561137, 1, [howick]], i54559067, 1, [howick]],  
[54559701, 1, [howick]], [54561795, 1, [howick]] .. . . .

4.4 An individual member of the list, such as i54559551, 1, [howick]], is made up of three components; the unique code for the tile, a measure of the mass of the association (in this case the number of residential delivery points (though in the case of a land cover type it would be the area of the cover within the tile).

4.5 To form derived areal units it is necessary to add morphology to mereology - to understand which tiles might be contiguous. As the question of contiguity depends upon granularity (any solid having spaces between its atoms if viewed at a sufficiently large scale), the assessment of contiguity involves generalization. For the present project this is achieved in three steps. The first is actually part of the intrinsic numbering system for the tiles, which gives their quadcodes (e.g. 54559551 above) - i.e. identifiers which allow them to be located in a 'quadtree' so that tiles that are close to each other are likely to have similar codes.

4.6 The second step might be thought of as generalizing the quadtree, and the third is finding contiguous tiles in the generalized tree. These matters are considered in more detail in Appendix 3. For the purpose of the present discussion it is sufficient to appreciate that each of the quadcodes can be expressed in base four, so that for example 54559551 becomes

[3, 1, 0, 0, 0, 2, 0, 0, 3, 0, 3, 3, 3]

- 4.7 This can be converted in turn to a tesseral representation and tesseral arithmetic can be used to determine which tiles are contiguous (even when the tiles are of different sizes). By generalizing (in this case to the 800m scale) and identifying contiguous cells and flagging codes that refer to contiguous blocks with a unique code it is possible to build a set of contiguous area represented by an augmented form:

[(11, [0,2, 3, 1,3,3,0,2,3, 11, 1, [generalized]], [12, [3, 1,0,0,0,2,0,0, 1, 1], 1, [generalized]], [13, [3, 1, 0, 0, G, 2, 0, 0, 3, O], I, [generalized]], [12, [3, 1, 0, 0, 0, 2, 0, 0, 3, 3], 1, [generalized]], [12, [3, 1, 0, 0, 0, 2, 0, 1, 0, 0], 1, [generalized]], [12, [3, 1, 0, 0, 0,2,0, 1, 0,2], 1, [generalized]], [12, [3, 1, 0,0, 0,2, 0, 1,2,0], 1, [generalized]], [12, [3, 1, 0, 0, 0,2, 0, 1,2, 2], 1, [generalized]], [12, [3, 1, 0, 0, 0, 2, 0, 2, 1, 11, 1, [generalized]], [12, [3, 1, 0, 0, 0, 2, 0, 3, 0, 0], 3, [generalized]]]

- 4.8 In this example, there are **two** contiguous areas (numbered here 11 and 12) which correspond to two separate settlements- one in the South West of England and one in the North East).
- 4.9 It is important to appreciate that this method of finding contiguous areas of concern by building and generalizing quadtrees, defining dense spaces and assessing contiguity provides a method of general significance. While it might be considered merely unfortunate that the settlement name Howick is not unique, generally there is no reason why the 'black tiles' identified by a particular substantive query should form a single geographic cluster. The general case is that the tiles identified may be arranged into lone or more clusters (representing different localities with a 'need' for farm diversification).

## 5 Impingement

- 5.1 In the discussion so far it has been assumed that an arbitrarily complex enquiry will yield a list of (black) tiles, on the basis of the objects that impinge upon the tiles. The term 'impinge', as noted above, should be taken to include, but not be limited to, spatial coincidence. Thus a list of dwellings or residential delivery points may be associated with the county of Cornwall because they lie within its bounds.
- 5.2 Royal Mail might of course actually direct mail through a post-town in a neighbouring county, and this might appear in the postal address, providing a first example of an association which is not about spatial overlap. For the purpose of the approach outlined in this document, it is important that the manner in which an object may impinge on a tile is expressed very broadly. Two examples might serve to illustrate the range

- of relationships which could be included within the term 'impinge' and also illustrate their pertinence.
- 5.3 First, consider the prospect of job losses in a particular industry. The industry is an object which might be thought of as including firms, which in turn might be thought of as occupying establishments which are workplaces to which individual employees travel. Industry, firm, establishment and worker are all objects which impinge (directly or indirectly) upon the tile where the establishment is located. The worker also has a home address which itself obviously impinges upon a particular tile.
  - 5.4 It is also clear, however, and substantively important that potential job cuts will (through this structure) impinge upon the tile where the worker's dwelling is located. For the purposes of this study **OA-OA** work travel data from the 2001 Census is used to establish such patterns. To illustrate, consider a firm e.g. Littlewoods which amongst its portfolio of establishments holds one with a unit postcode W. Within the Prolog framework, the fact that W is in a particular **OA** (say O) is held. Separate facts express the knowledge that n workers travel from each possible origin **OA** (say H) to W and workers resident in a particular postcode H , constitute a proportion p of all workers.
  - 5.5 A second illustration concerns diffuse pollution. Various farming practices in both the arable and pastoral sectors result in pollution of watercourses into which the affected tracts of land drain. Clearly these practices impinge upon each tile within the area of activity. They also have an impact upon the downstream river network, and so upon all the tiles on which this network object impinges. Clearly, this could be taken further by looking at the abstraction of water. For present purposes, however, it is important simply to note that within the logic-programming framework, the relationship between any objects may serve to form part of a logical chain which may specify which attributes of which objects impinge upon a particular tile.

## **6 Concluding Remarks**

- 6.1 Hopefully the foregoing paragraphs will serve to illustrate the manner in which existing data can be used to define the objects which form the content of policy talk, and to draw out value. The focus has been on specifying relationships between the objects of policy talk, the geographic areas on which they impinge and stocks of available data. The overall goal has been to prototype tools which allow greater flexibility in specifying problems and geographic objects and hence allow for more fruitful exploitation of existing data.

- 6.2 It should be understood that given the required flexibility and the evident disjunction of data and policy concerns, it is extremely unlikely that the forms of analysis discussed should be treated as entirely automated procedures. Clearly, the specification of logical linkages must imply a degree of understanding (and a degree of tenacity) on behalf of the user. This is, however, far from suggesting that the sorts of functionality discussed in this document should not be available to DEFRA stakeholders either directly or indirectly. Only with such functionality is it possible to realise the intention of effectively linking evidence and policy.

### **Reference**

R. Casati, B. Smith, A. C. Varzi, *Ontological Tools for Geographic Representation*, in: *Formal Ontology in Information Systems*, edited by N. Guarino, pp. 77-85, IOS Press, 1998