

# **“Gis a job”: What use Geographical Information Systems in Spatial Economics?<sup>1</sup>**

Henry Overman (LSE and SERC)<sup>2</sup>

*Abstract:* Geographical Information Systems (GIS) are used for inputting, storing, managing, analysing and mapping spatial data. This article argues that each of these functions can help researchers interested in spatial economics. In addition, GIS provide new data which is both interesting in its own right, but also as a source of exogenous variation.

---

<sup>1</sup> See [http://en.wikipedia.org/wiki/Yosser\\_Hughes](http://en.wikipedia.org/wiki/Yosser_Hughes).

<sup>2</sup> Department of Geography and Environment, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom (email: [h.g.overman@lse.ac.uk](mailto:h.g.overman@lse.ac.uk); website: <http://cep.lse.ac.uk/~overman>). Also affiliated with the Centre for Economic Performance and the Spatial Economics Research Centre at the London School of Economics.

## **Introduction**

Geographical Information Systems (GIS) are widely used in business, government and a growing number of academic disciplines. They are clearly “useful”. As an economist interested in spatial issues, however, it has not always been obvious in what way they might be useful to me (and others similarly interested). This article is an attempt to provide a partial answer to that question.

It does so in several steps. I start by providing a very brief description of GIS. I then discuss how GIS helps deal with spatial data. Dealing with spatial data can be complicated. I try to highlight some of the problems, but also flag some things that GIS experts worry a lot about, but that do not seem to matter much for applications in spatial economics (especially when GIS analysis is used as an input in to further statistical or econometric analysis). Standard GIS textbooks would spend several hundred pages carefully discussing these issues; I will cover them in around 2,000 words. GIS experts will want to skip these sections, if only to keep their blood pressure low.

The primer covers a number of ways in which GIS can help handle spatial data. A final section turns to the question of how it can help increase our understanding of socio-economic processes. In short, I argue that in addition to the role of GIS in facilitating research with spatial data it also helps avoid arbitrary discretisation, provides interesting new sources of data and of exogenous variation that allows the construction of innovative instrumental variables. It is less clear to me how some “cutting-edge” advances will help solve the deep problems of spatial economics and the article finishes with brief consideration of those issues.

## **Geographical Information Systems**

Geographical Information Systems are used for inputting, storing, managing, analysing and mapping spatial data. Of course, a wide range of software can provide similar functions for quantitative data. It is the explicit focus on the geographical, or spatial, element that makes GIS unique.

Traditionally, spatial data has come from two main sources. Ground survey (what goes on at a particular location) and census (what does the distribution look like across locations). These sampling and census data will be familiar to researchers interested in spatial economics. Less familiar, perhaps, are data collected using sensing devices located remotely (possibly far from both the object being studied and the data collector). Examples include aerial photography and, increasingly, data from satellites (either locational or radiometric). The latter are often referred to as providing remote sensing data, although sometimes that term is applied more generally to any data collected indirectly. Remote sensing typically provides large amounts of data on the earth's surface. These data are of inherent interest but, as discussed below, may be of particular interest to researchers seeking sources of exogenous variation to use as instruments.

The means of inputting data in to the GIS depends on the source. One can move from analog maps to digital data by either digitizing (manually tracing over the map) or scanning in the map. Both of these processes usually require a considerable amount

of skill and most researchers will want to outsource these tasks (on the basis of either absolute or comparative advantage). I would make a similar comment with respect to processing remote sensing data from satellites.

One proviso, however, is that it can be very important to understand the sources of measurement error that can occur during the inputting process. Some of these are specific to the way that GIS handles data (e.g. a polygon representing a lake is open, but should be closed; lines on a transport network should intersect but do not) and some are so general as to be not worth considering in detail here (e.g. forgetting to input a line of data). Working on the assumption that your local GIS expert has been careful and sorted the GIS specific problems for you, the more interesting issues arise because social scientists may sometimes think very differently about measurement error and the problems that it may create for subsequent analysis.

Let me give one example from personal experience. In Burchfield et al (2005) we were trying to characterise urban sprawl and what might cause differences across metropolitan areas using remote-sensed land use data for circa 8.7 billion 30m by 30m cells covering the US. Remote sensed land-use data may be either “leaves-on” or “leaves-off”. If you take a leaves-on land use image of, say, Massachusetts then you miss all those urban land use features sitting under trees. Ever-green trees could give you the same problem all year round. Whether you need to worry about this depends on what you are trying to do with the data. When trying to explain the causes of sprawl we used “leaves-off” data focussing on metropolitan areas to minimise measurement error. In our regressions, we then used region dummies as well as latitude and longitude variables to allow for the fact that the extent of this problem might vary systematically across space. Given that our results were robust to the inclusion of these control variables we were reasonably comfortable that systematic measurement error (i.e. in some way correlated with our explanatory variables) was unlikely to be biasing our results. A surprising number of people that we presented these results to just could not get past the fact that the remote sensed data measured land use *with error* despite the fact that the dummies and controls should to a large extent mitigate the impact of this error on our results. We also received similar complaints about the fact that the 30m by 30m resolution data we used was not as accurate as, say, land-use parcel data available in particular counties of the US. We thought the trade-off of in terms of slightly less accurate data for the entire US versus more accurate data for one particular county worth making to increase our understanding of the factors determining urban sprawl across the US. It is clear that a number of researchers do not agree but again, what is surprising, is that the discussion is often about the existence of measurement error *per se* rather than the impact that this might have on results.

Before leaving issues of inputting and measurement error, one final point on the use of Global Positioning Systems to reduce location measurement error in household surveys (particularly in developing countries). It is clear that this could be a major issue when analysing, say, the determinants and consequences of access to public services. Gibson and McKenzie (2007) discuss use of GPS in household surveys and show that self reported distances can correlate rather badly with actual distance based on GPS measurement. Interestingly, they also suggest that straight line distances based on point-to-point GPS measurement are very highly correlated with much more complex calculations based on actual transport networks. This last point echoes the

finding of Combes and Lafourcade (2005) that measures of transport costs based on straight line distances perform reasonably well for cross-section data (but rather badly for time series changes). All of this also raises the question of which measure of distance (actual or perceived) matter for individual behaviours.

These examples serve to highlight the simple points that (i) as for all data, sources of measurement error matter for GIS and (ii) social scientists can bring a better understanding of the *consequences* of measurement error in GIS data providing we understand the sources of those errors. I draw similar conclusions from thinking about the way GIS helps store and manage data from a variety of sources. It is to this issue that I now turn.

Beyond the use of GIS to map data (of which a little more a little later), it appears that growing numbers of social scientists are beginning to use GIS to reconcile spatial data from different sources to create new data sets (this tends to be referred to as overlaying).<sup>3</sup> At its simplest, this involves using GIS to merge different socio-economic data for the same spatial units. Non-spatial software can handle this easily, so let us turn immediately to the more interesting issue of the use of GIS to merge socio-economic data recorded for *different* spatial units. For example, household data from census tracts with firm data for post or zip codes. Of course, many social scientists already use concordances to map data from one set of spatial units to another. GIS is particularly useful when such concordances are not readily available or, more interestingly, when such concordances are difficult to construct using standard econometric or data management software because the data are recorded using non-nested spatial units.

To understand how GIS helps solve these kind of problems and to consider the issues that arise, we need to briefly consider how data is usually represented in GIS. That is, what kind of geographical data models GIS uses to store data. There are two common formats: raster and vector. Raster format organises spatial data by assigning values to each cell on a regular grid (this is usually square, but does not need to be providing it provides coverage of the area under consideration). In contrast, vector format assigns values to irregular polygons and then provides coordinate data on the location of these polygons. Most GIS software will provide a variety of tools to move between the different representations and to merge “layers” of different data recorded using either format. The methods used are often rather intuitive but, as always, the detail matters. As with the pre-coded routines that comes with standard econometric packages one should have (or be employing someone who has) a reasonable grasp of how these transformations occur.

The standard references cover the issues concerned in some depth and the reader is referred there for details. As with inputting data, I will not attempt to provide a systematic discussion of the issues concerned. Some of these are, once again, specific to the way that GIS handles data. For example, because areas in a vector data set may be defined separately, the common boundaries for two neighbouring areas may not lie on top of one another. If one is just merging in data that are all recorded for the same spatial units this will not create a problem. But if you are merging in data for

---

<sup>3</sup> GIS specialists tend to treat this as an issue of analysis but I think most social scientists would describe this in their “data section” as an input to the spatial analysis to follow.

different spatial units it can do. For example, London's electoral wards can be divided into those in outer London and those in inner London. Imagine defining the boundary between inner and outer London on the basis of inner London wards. Now ask the GIS to find *all* wards that have some area inside that boundary. Because common boundaries need not lie directly on top of one another some outer London wards that are contiguous to inner London wards will end up with small areas inside the inner London boundary just defined. If one then tries to overlay inner London data on the ward map of London using the ward-defined inner London boundary these neighbouring outer London wards get the inner London value. Of course, it is easy to envisage a non-spatial solution to this kind of aggregation problem because the inner/outer London distinction is an exact aggregation of London electoral wards. But GIS is very useful in situations where that is not necessarily the case, so it is useful to understand how these misclassification errors arise. In this example, identifying inner London wards as those with "ward centroids within the boundary" rather than "any area of the ward inside the boundary" would almost certainly fix the problem.

Even raster data that may appear to be reported for an identical grid can suffer from these kind of problems as, for example, small variations in satellite positioning may slightly change the gridding of remotely sensed data. The solution to these kind of problems is complex, but they are sufficiently pervasive that, say, first differencing 30m by 30m LandSat land use pixel data from 1990 and 2000 to talk about individual pixel level changes comes with a very significant health warning.

Aside from these technical problems, there are interesting conceptual questions about what rule you should use to assign value data when merging vector data layers that are recorded for different (non-nested) spatial units. Similar considerations will almost certainly apply when trying to merge raster data to vector data (unless there are some very "boxy" looking spatial units in the vector data). The correct answer, of course, depends on the underlying spatial distribution of the phenomena for which one has measured attributes. For example, if a phenomena is uniformly distributed across space then area weighting can be used to move between spatial units. Non-uniform distributions may call for weighting by some other spatial characteristic (although this characteristic must then, be available at smaller spatial scales to allow for weighting). See, for example, Huby, Owen and Cinderby (2007) for discussion. These decision rules open up interesting areas of overlap between spatial statistical techniques such as kriging and the rules that GIS uses to deal with non-overlapping spatial units. There is literature on this, particularly concerning the application of GIS to continuous surface data. However, it would be fair to say that the literature is fairly technical and may not readily accessible to more than a small number of specialists. That said, it is certainly useful to be aware of the problem, particularly because it may not necessarily be clear what rule a particular GIS is using when it overlays data. This also raises the question of the statistical properties of matched survey (rather than census) data. The effects of such matching have recently been considered for non-spatial data when matching data for the same observational unit from different survey data sources. I am not aware of much discussion of the additional issues this would raise in circumstances where the underlying units of observation are not identical as is the case for non-nested spatial data of the kind discussed here.

The most powerful aspect of GIS is arguably its ability to quickly analyse spatial data. These tools have been little used in spatial economics outside of a relatively small, technically proficient group of users. The increasing accessibility of GIS software and growing interest in spatial issues looks set to change this. I provide a partial review below drawing, where possible, on specific examples from the literature.

GIS can be used to identify observations by both characteristics and location and then to perform simple statistical operations (e.g. counts). The two most common examples in spatial economics are probably the study of land use and the hedonic analysis of house prices. For example, the 30m by 30m land use raster data we used in Burchfield et al (2005) categorizes circa 8.7 billion cells into one of 21 land cover classes. We then use this data to ask questions, say, about the amount of developed land in each state or Metropolitan Statistical Area. In hedonic analysis GIS can be used to identify and characterise properties of the residence and the parcel on which it sits as well as wider neighbourhood characteristics derived by merging the point data locating houses to other socio-economic data (see Bateman, Jones, Lovett, Lake and Day (2002) for a review).

Locating and characterising observations is the most basic form of analysis in GIS. However, much more complex analysis is possible. Given that GIS data are spatial, a natural use of GIS is for measurement. For example, GIS can be used to measure the length of lines, the perimeter or area of polygons and the distance between observations or between observations and other features of interest. These distances could be physical distances, network distances (e.g. along a transport network) or involve some more general concept of social distance. GIS can also use information on absolute barriers (e.g. impassable rivers) to calculate shortest path-distance. Additional information on movement costs, be they continuous (e.g. gradient) or discrete (a border crossing), can be incorporated to calculate least cost surfaces or paths (along a network). GIS can also be used to measure shape (e.g. to capture sinuosity by taking the ratio of straight line to actual distance).

I am not familiar with applications of sinuosity outside of environmental/ecological research, but there may be some. Cheshire and Sheppard (1995) have used measures of polygon shape to characterise back gardens and their impact on housing values in a hedonic analysis. Least cost path analysis has been widely used in the transport economics literature to calculate transport costs between locations. These methods are now being increasingly applied in other fields. For example, Donaldson (2009) uses least cost path analysis to calculate transport costs in his fascinating paper studying the economic impact of India's vast rail network. Faber (2009) goes one step further and uses data on impedance values of gradients, wetlands etc to calculate least cost routes between major Chinese cities and then a spanning tree algorithm to construct a the least cost network. He then uses this network as an instrument for the actual Chinese highway network in a study of the economic impact of roads. The problem that this addresses is a well understood one – how do we identify the causal impact of roads on local economies when local economic outcomes help determine road placement? For places that lie in between the major Chinese cities (the nodes on the network) being close to the path followed by the least cost network should increase the probability of being on the actual network but should not affect economic outcomes otherwise. It thus provides a valid instrument. Similar, in spirit, are the identification strategies used by, for example, Baum-Snow (200x), Michaels (200x),

Durant and Turner (200x) and Donaldson (2009) all of whom rely on historical plans (e.g. for the interstate highway system) or historical networks as instruments for the actual network. These papers all use GIS to construct their instrument but Faber (2009) is the first paper that I know of to use the alternative strategy of building his instrument using GIS analysis to provide the lowest cost network. Of course, such techniques have been extensively used in transport economics and planning, what is interesting here is the use of these techniques to construct instruments to help identify the causal effect of roads on spatial economic outcomes. I discuss this further below.

GIS experts will sometimes distinguish between the use of GIS to measure distance of observations from other features and its use to calculate the distance between observations. While the former could be regarded as a specific example of the use of GIS for measurement, the latter provide one specific example of the use of GIS to understand spatial arrangement. Hedonic analysis provides plenty of examples of the use of GIS to measure distance from observations to other features. To take just one example, Gibbons and Machin (2005) use GIS to measure the proximity of properties to rivers, coasts, woodlands, roads, railway lines and airports in their study valuing rail access. Turning to the broader issue of spatial arrangement there are a number of potential areas of overlap between GIS and spatial economics which have only just begun to be explored. I consider these issues next.

Researchers in biology and biomedical sciences already make extensive use of observation to observation distance in their statistical modelling of spatial point patterns (see Diggle, 2003). If distance affects the strength of the interaction between observations (e.g. the chance that I catch a disease from you) then knowing the relative position of observations can help understand outcomes (e.g. incidence of a disease). Examples abound in spatial economics. For example, in models of spatial competition the intensity of competition faced by firms may depend on the distance between them and rival firms. In models of matching, the chance of, and payoff from, a match can depend on the average distance between individuals. Observation to observation distances are also useful when we want to assess whether there are systematic patterns evidenced in individual location choices. For example, studies of localisation assess whether firms in a specific industry tend to be spatially concentrated (or dispersed) relative to overall economic activity. If they are, then observation to observation distances for firms in this industry will be less than for firms randomly chosen from the economy at large (or more than if the industry is dispersed). More generally, the distribution of observation to observation distances (or statistics based on that distribution) should allow us to assess both the existence and extent of any systematic departures from random location. The increased availability of geo-referenced economic data should see these distance based techniques become more common in spatial economics.

In the context of this article I should note, however, that it is not completely clear to me whether this should always involve the use of GIS to construct these distances. Simple one-off calculations of distance from observation to feature (e.g. for use in a subsequent hedonic regression) are fairly easy to implement in GIS. There is also a considerable time saving to be had as soon as the calculations become more complicated. GIS software is, for example, quick at finding nearest neighbours. Its comparative advantage becomes greater as the number of observations increases. This is because the brute-force approach of, for example, taking the distance between all

observations and identifying the minimum involves a rapidly increasing number of calculations as the number of observations grows. In contrast, because GIS uses algorithms that reduce the rate of increase in the number of calculations, it becomes increasingly efficient as the number of observations grows. Of course, one could incorporate these algorithms in to non-GIS routines but the sunk costs of doing so may well exceed those of familiarising oneself with an off the shelf GIS. Personal experience suggests that this may not be the case for point to point analysis when bootstrapping is nearly always needed to calculate statistical significance. For example, in Duranton and Overman (2005) we used the distribution of distance between firms to assess the location patterns of around 250 UK industries with, on average, 700 plants per industry. Bootstrapping the global and local confidence intervals involved 1000 simulations per industry. We then used non-parametric kernel techniques to estimate the distribution of bilateral distances for each simulation. We did this in gauss using a non-parametric kernel routine taken from an existing library with a simple C+ routine providing a first pass binning of the data for speed. In principle, one could write batch files (i.e. where the user writes a sequence of commands in a file that the computer implements one by one) to achieve something similar in GIS. This involves fairly large fixed costs in terms of both purchasing software and learning how to implement the relevant procedures. Personally, I would not know how to automate such a procedure in GIS and it may well be that many spatial economists would similarly find the non-GIS route a more efficient way of dealing with the problem given their background. I am increasingly convinced, however, that the sunk costs are worth incurring in many simpler situations where GIS calculations should be either more accurate or considerably faster than short cuts implemented using non-spatial software.

A good example of this is the use of GIS to define neighbourhoods (or “buffers”) around objects. For example in Burchfield et al (2005) we use GIS to calculate the percentage of the urban fringe - defined as a 20 kilometer buffer around existing development - that lies above water yielding aquifers. The use of such buffers is in its infancy in socio-economic applications but has the potential to be very useful because it reduces the need for research to rely on arbitrary discretisations of the study area of interest. For example, in Duranton et al (2006) we study the impact of local taxation on firms. To do this, we build on research by Holmes (200x) who controls for unobservables by using counties on either side of state boundaries with different labour laws. That is, the identifying assumption is that firms in these neighbouring counties only differ in terms of labour laws and otherwise face similar environments. In our work, because we have geo-referenced data, we do not have to rely on arbitrary county boundaries to control for unobservables. Instead we simply identify all firms that are located within a given *distance* of the jurisdictional boundaries that separate the authorities who set the local tax. Whether this is useful will depend on the degree to which boundaries are arbitrary with respect to the phenomena under study and the extent to which there is significant variation in unobserved conditions within spatial units defined by the set of boundaries. It certainly provides a useful step forward when boundaries are arbitrary and the degree of unobserved variation within spatial units is large.

So far, I have covered the use of GIS for inputting, storing, managing and analysing spatial data. I have saved the most frequent application in spatial economics to last – the use of GIS to visualise or map economic data with a spatial component. I do not



have much to say on this. Most entry level courses in econometrics begin with a plea to “plot the data” at an early stage of analysis to help identify trends, outliers etc. Much the same could be said of the role of mapping spatial data and GIS provides a simple and efficient way to do this. Mapping raises a number of issues to do with the appropriate representation of spatial data. Some of these have clear non-spatial analogues. For example, the appropriate binning of data for a choropleth map involves similar issues as does binning for a histogram. Some are more specific, for example, how to overlay a number of features occurring at the same location? Standard GIS textbooks provide further discussion.

### **Can Geographical Information Systems help increase our understanding of the spatial economy?**

So far I have talked a lot about what GIS can do in a rather instrumental way (i.e. what it can do). For the final part of this article I want to focus on the extent to which it can help us increase our understanding of socio-economic phenomena. In part this provides a summary of the discussion above, but I also want to use it to consider the way in which GIS helps introduce us to new data and to new sources of exogenous variation.

I have already considered the ways in which GIS can reduce measurement error, particularly with respect to the more precise location of observations. This issue has received little attention in the literature, perhaps because there was relatively little, hitherto, that researchers could do about it. As discussed above, the increasing use of GIS and the availability of geo-referenced data look set to change this.

The somewhat related issue of the arbitrary discretisation of continuous space has received considerably more attention. This is particularly the case in studies which focus on the level or growth rate of economic outcomes for spatial units (be they neighbourhoods, cities or regions). Two common solutions have been adopted. Either researchers have sought to identify “functional areas” using a variety of criteria (e.g. commuting flows to give travel to work areas) or else they have presented results at a variety of spatial scales. Similar problems and solutions have been discussed in the literature concerned with the spatial patterns of location of particular activities and other areas of interest. Once again, hitherto, there was little more that researchers could do about this problem. The increasing use of GIS to allow reconciliation of data for different, possibly non-nested, spatial units will help address this problem too. In some circumstances geo-referenced data should allow researchers to circumvent the problem altogether by switching to continuous space. See Duranton and Overman (2005) for a specific example and further discussion. My feeling, however, is that the solution of switching to continuous space is more readily applicable in some situations than in others. For example, if we are concerned with structuring our analysis using general equilibrium constraints then it can be difficult to think how to impose these in continuous space. There have, of course, been theoretical advances along these lines so empirical implementation is not inconceivable, but I don’t think we are there yet. I will consider the role of theory a little more below.

As researchers become more familiar with using GIS to integrate data from different sources, they will be increasingly exposed to new sources and types of data that can help increase our understanding of spatial economic phenomena. I have already

discussed the fact that remote sensing data from either satellite or aerial photography can provide a vast amount of data on the earth's surface. Digitised geological maps provide a further source of information. These data have already seen extensive use in natural and environmental resource management, as well as in agricultural economics. Outside of these narrower fields, data on land cover and land use (i.e., the physical features that cover the land and what those features are used for), soil type, geological and landscape features, elevation and climate are increasingly being used to construct explanatory variables to help increase our understanding of different features of the economic landscape. Consider a few examples. At least since Rosen (1974) urban economists have been interested in the impact of climate on city population. In contrast to earlier studies, however, Rappoport's (2004) research is able to use meteorological data that comes from 6,000 meteorological stations and covers 20 winter, summer and precipitation variables. GIS analysis by the Spatial Climate Analysis Surface at Oregon State University applied to this meteorological data allows the construction of weather variables for a 2 kilometre grid covering the continental U.S. Deschenes and Greenstone combine soil quality data for 800,000 sites combined with 4km by 4km grided precipitation and temperature data to provide an estimate of the impact of climate change on agricultural output. In a very different context, Nunn and Puga calculate the ruggedness of different countries using a global elevation data set whose underlying spatial units are 1km by 1km squares. They then use this data to study the impact of ruggedness in Africa with the startling finding that the indirect effect (rugged terrain protected against slave traders) outweighs the direct effect (ruggedness hinders trade and productivity). It is hard to see how any of these exercises would have been feasible without the availability of GIS data and analysis.

New data, particularly on features of the earth's surface are also great sources of exogenous variation. As a result, researchers are increasingly using GIS data to construct innovative instruments to help identify causal effects in a range of different literatures. Some examples should help to make this idea concrete.

Hoxby (2000) is interested in whether competition among public schools improves outcomes. In systems, such as the US, where school districts have strong control over schools (rather than individual schools making their own choices) competition amongst public schools should get stronger as the number of school districts increases. That is, cities with more school districts should have better public schools and less private schooling. The problem in examining this hypothesis is that, conditional on city size, better public schools and less private schools should imply more school districts (assuming school districts are roughly equally sized across cities). Because the number of districts is endogenous to public school quality we need an instrument that determines the supply of school districts but that is independent of the local public school quality. Hoxby argues that the number of streams in a metropolitan area provides such an instrument because cities with a large number of streams end up with more school districts for reasons unrelated to school quality.<sup>4</sup>

On the subject of streams, Duflo and Pande (2007) use GIS to help study the productivity and distributional impact of large irrigation dams in India. The problem

---

<sup>4</sup> This paper provides a well known example of the strategy, although not of the use of GIS as her work is based on the study of detailed paper maps.

with studying this impact is that the citing of dams is not random so that regions with and without dams are likely to differ along other dimensions (e.g. agricultural productivity). However, it turns out the gradient at which a river flows affects the ease of dam construction. Irrigation dams work best with low (but nonzero) river gradient. Hydro electric dams work best with steep river gradient. They use topographic data from the same source as Nunn and Puga (2009) to characterise elevation and gradient and to construct instruments that are positively correlated with the likelihood of getting an irrigation dam but otherwise independent of agricultural productivity.

For our final couple of examples, consider the use of GIS to construct instruments that can be used to help identify the causal impact of density on productivity. Again, the problem itself has long been recognised. High density may cause high productivity, but high productivity for some other reason will in turn attract firms leading to high density (rather than vice-versa). Once again, what we need is something correlated with density but independent of productivity. Rosenthal and Strange (2005) provide a neat example by noting that the density of employment will be partly determined by the height of buildings in a location. Building height, in turn, is partly dependent on the underlying geology of the site. Given that, outside of agriculture, geology should not determine wages directly the underlying geology can be used as an instrument. They use GIS data on the type of underlying bedrock, seismic and landslip hazard as instruments for the density of employment in their regressions of wages on employment density. Combes et al (2009) use a similar idea and data from the European Soil Database for 1 km per 1 km cells to construct a number of instruments describing the mineralogy of the sub and top-soils, the nature of the dominant parent material at broad and detailed level, seven other characteristics of the soil such as water capacity as well as the ruggedness of the terrain.

These examples are clearly not extensive. However, they do serve to give some flavour of the emerging research and I think all of this suggests a potentially important role in future work for GIS data as a component in novel instrumentation strategies. I think there will also be a place for GIS to play an increasing role in the treatment effects literature. Linden and Rockoff provide a nice example where they exploit both the timing of move-in and the *exact* location of sex offenders to improve on existing estimates of the impact of crime risks on property values.

To many GIS specialists, this might all sound rather bread and butter. Where are all the cutting-edge advances? Here, I think my own disciplinary background as an economist strongly colours my thinking. For me, the deep issues in empirical spatial economics are unlikely to be addressed by these advances. For example, what role should theory play in structuring our empirical analysis? Tiebout models of residential segregation are useful in structuring the analysis of residential sorting across neighbourhoods (e.g. as demonstrated by Epple and coauthors). What role could they then play in furthering the empirical study of the production structure of regions (e.g. as in Hanson (199x), Combes and Lafource (200x), Mion (200x)). Or is the latter situation too complex to be amenable to full structural modelling and thus better suited to the kind of mixed estimation/calibration exercises that appear to be increasingly popular in, say, the international trade literature (e.g. Eaton and Kramatz (200x))? What role should theory play in achieving identification of causal effects? Instead of theory, should the emphasis be on the treatment effects literature that is

increasingly popular in the labour literature? Or on instrumental variables? Or on setting up experiments involving researcher generated exogenous variation as seems to be increasingly popular in, say, the development literature?

A quick look at the cutting edge of GIS analysis and modelling suggests that it is grappling with a very different set of issues. It so happens that as I was writing this piece I received an invitation to the Third International Cartographic Association Workshop on Geospatial Analysis and Modelling. Here is a quick run down of the list of highlighted topics: Cellular automata and agent based modelling; Visual analytical tools for environmental and urban systems; Analysis of human movement data; Spatio-temporal data mining; Hierarchies, scaling and fractal structure of geographic patterns; Modeling vehicle dynamics and crowd behaviour; Patterns of human spatial behavior and migration; Urban high-resolution morphology; Small world modeling and spatial interactions. If one was to attend, one would expect to see lots of very cool graphics, some very neat simulations and some serious spatial statistical modelling. This would have strong synergies with some areas of regional science that have increasingly relied on models from physics and biology to structure their analysis. Most spatial economists find it difficult to engage with that agenda (see, for example, Overman 2009) and our priorities, as outlined above, lie elsewhere. Economists are uncomfortable with predictions, but for what it is worth, I see myself sticking to the GIS bread and butter diet for some time to come.

## References (incomplete)

- Bateman, I. J. and A. A. Lovett (1998), 'Using Geographical Information Systems (GIS) and Large Area Databases to Predict Yield Class: A Study of Sitka Spruce in Wales', *Forestry* 71 (2), 147–168.
- Bateman, I., A. Jones, A. Lovett, I. Lake and B. Day (2002) *Applying Geographical Information Systems (GIS) to Environmental and Resource Economics*. *Environmental and Resource Economics* 22: 219–269.
- Burchfield, M., H. G. Overman, D. Puga and M. Turner (2005) *The Determinants of Sprawl: A Portrait From Space*. Forthcoming *Quarterly Journal of Economics*.
- Diggle, P. (2003) *Statistical Analysis of Spatial Point Patterns*, Arnold, London.
- Duranton, G. and H. G. Overman (2005) Testing for Localisation Using Micro-Geographic Data, *Review of Economic Studies*, 2005, 72(4), 1077-1106.
- Gibbons, S. and S. Machin (2005) Valuing Rail Access Using Transport Innovations, *Journal of Urban Economics*, 57: 148-169.
- Hoxby, C. (2000), "Does Competition Among Public Schools Benefit Students and Taxpayers?", *American Economic Review*, 90(5), 1209-38.
- Huby, Owen and Cinderby (2007) Reconciling soci-economic and environmental data in a GIS context: An example for Rural England.
- Longley, P., M. Goodchild, D. Maguire and D. Rhind (2005) *Geographical Information Systems and Science*, 2<sup>nd</sup> edition, John Wiley and Sons.
- Overman, H. G. and Y. Ioannides (2004) The Spatial Evolution of the US Urban System, *Journal of Economic Geography* 4 (2), 131-156.
- Rappaport, J. (2004), *Moving to Nice Weather*, Federal Reserve Bank of Kansas City Working Paper #03-07.
- Rosenthal, S. and W. Strange (2005), *The Attenuation of Agglomeration Economies: A Manhattan Skyline Approach*, Processed University of Toronto.