National Historical Geographical Information System as a tool for historical research: Population and railways in Wales, 1841-1911

Ian Gregory and Robert Schwartz

1. Introduction: Historical GIS and National Historical GIS databases

Historical Geographical Information Systems, or Historical GIS, has become a rapidly growing field within historical research (Gregory & Ell 2007; Gregory & Healey 2007; Knowles 2005). It is an inter-disciplinary field that involves taking GIS technology, devised in the fields of computer science and environmental management, and applying it to the study of history. A major impetus behind the growth of Historical GIS has been the significant investments made by a number of countries in National Historical GISs (NHGIS). Typically, these databases contain a country's census reports and other data for the nineteenth and twentieth centuries, including both the statistical data linked to specific administrative units and their geographic locations, together with the changing boundaries of those units. Using a traditional census database containing only statistical information, a researcher could search for *aspatial* patterns of variation and change across units unattached to their real geographic locations. Using an historical GIS, however, the researcher is now equipped to identify patterns of change that occur simultaneously over time and across geographic space. With historical GIS we can get closer to complexity of change and historical reality.

Among the countries that have built or are building NHGISs, the best developed include Great Britain (Gregory et al. 2002), the United States (Fitch & Ruggles 2003), and Belgium (De Moor & Wiedemann, 2003).[1]. Such systems are costly because they require not only the entry of census information but also the researching and encoding of administrative boundary changes through time, not to mention myriad challenges along the way.[2]

Is all the effort and money expended worth the cost? This paper offers a strong affirmative answer in two ways. Using the Great Britain Historical GIS as an example of highly developed national historical GISs, this paper describes how it was built and illustrates both the analytic approaches and the substantive results that historical GIS enables. Computer-assisted GIS is not the philosopher's stone of course. But it is capable making new contributions to our knowledge of the past, all the more so when combined with other tools in the historian/historical geographer's kit.

2. Building an NHGIS: Theme, space and time

[Figure 1: GIS data model]

Simply put, a GIS is a database with a difference. The difference lies in the specific connection made between information about a specific place and the precise spatial location of that place. In the terminology of GIS, we call the information about a place *attribute data* and the geographic reference, *spatial data.* Typically, spatial data are represented in one of four forms: a point, a line, a polygon (that represents a zone or area), or a grid cell in a graphic image. This structure is shown in figure 1. Historical GIS is concerned with taking data in this basic form, and using it to better understand the geographies of the past.

[Figure 2: A table from the census]

Aggregate census reports enumerate the population, its characteristics, and related information within a clearly defined administrative unit. Usual categories include age, sex, marital status, employment, household composition, housing conditions, and so forth, and the resulting summaries are published in tabular form as shown in figure 2. In this example, each row describes the attributes (counts, percentages, etc.) for a location specified by place-name or identifying code; each cell in the row contains a count of the people in a given category. The rows are usually arranged in alphabetic order. Unit B follows Unit A. However useful, a database of this kind lacks an important capability. What is missing is an accurate representation of the underlying geography of the data. Unit A's spatial relationship with Unit B remains unidentified so whether they are neighbours, whether they are close or distant is not recorded or readily studied.

This missing geography matters because location, proximity, or distance may help account for the similarity or difference in patterns of employment, age/sec ratios, household structures, and so on. Historical GIS remedies this situation by incorporating spatial and attribute data. It thus provides the crucial capability to undertake temporal and geographic analysis.

The major challenge for a national historical GIS is therefore to re-unite the tabular data with the administrative units to which they refer. The difficulty in doing this is that maps are not necessarily available for the same dates as censuses, and that boundaries change frequently over time. This required the building of a GIS database of changing administrative boundaries in a form that could easily be linked to tabular lists of census and other data. The three types of administrative units that were required for England & Wales were the 635 registration districts from 1841-1911, the 1,500 local government districts from 1911-1973 (which consists of county, municipal and metropolitan boroughs, and urban and rural districts), and underlying these, the 15,000 parishes from 1871-1973.[3] During this period most census data were published at district level but the parishes were felt important as information on total population was available at this level which provides far more spatial detail than districts. The project did not attempt to come up to date because there were significant boundary reforms in England & Wales in 1973 and most recent data were already available in GIS form.

[Figure 3: Maps]

Figure 3 shows a sample of the maps used to create the system. Around 1910 a series of county maps were published at 2 miles-to-the-inch scale (approximately 1:125,000) that show all three sets of administrative units and are free from Ordnance Survey copyright. These were used for our main digitising. The series was reproduced around once every ten years after this although the scale changed to 1:100,000 after the Second World War. Prior to 1910 sources are much scarcer. A 4 miles-to-the-inch scale map (approximately 1:250,000) was available in 1888 and the 1851 census published a series of sketch maps but more detailed information could only be obtained from series of 1 inch maps in the National Archives from 1891 and around 1870 (Gregory et al. 2002).

To make this a full database of changing boundaries required additional information on where and when boundaries changed. This was available in tabular form in the *Registrar General's Decennial Supplements* where simple tables list the administrative unit that lost territory, the unit that gained it, a brief description of the area affected, and a precise date when the change occurred. Combining the spatial data from the map series with the temporal data from the boundary change tables provided enough information to re-create all of the major administrative units of the past together with an exact date at which their boundaries changed. Once stored in a GIS (ArcInfo software) a user can specify the geographic unit of analysis, the exact date of the unit's boundaries, and the region in which it was located. Using custom-written macros, the system will extract the requested boundaries in the form of an ArcInfo coverage or an ArcGIS shapefile.

As the digital boundaries were created, another project team entered information from the decennial census and vital registration tables, creating a database of over 20 million raw data values. This database contains a place-name for each row of data. As the GIS database contains a place-name for each administrative unit that it holds, linking the two could be done by joining on place-name. Differences in spellings between the two databases can be handled using place-name gazetteers.

[Figure 4: Population density in Wales]

This structure means that creating Choropleth maps such as the one shown in figure 4 becomes a quick and straight-forward process. Parish boundaries for Wales in 1911 have been extracted from the GIS and linked to population data from the 1911 census to give a map of population density. The entire process of creating a map takes only a few minutes, enabling the researcher to experiment with different class intervals, shading schemes and captioning and thus to explore the data in multiple ways and greater depth.

Choropleth maps provide a useful way of identifying and analysing geographical patterns within the data, which in this case consists of demographic data for 1,250 parishes. The map shows that the valleys area of south Wales had particularly high population densities, a pattern associated with the concentration of coal mining and heavy industry in that region. In north-eastern Wales, parish population densities were also high. Thanks to the detail preserved in the parish-level data, the map reveals a ribbon of relatively high density areas along the coast as well as a broad swath of low density areas in the inland parishes of Central Wales. The relatively high densities of some towns compared to their rural hinterlands also stands out when the data is viewed at the parish level.

Creating the GBHGIS took around six years and over £500,000 of funding, most of which was spent on researching and digitising changing boundaries. The resulting system integrates map-based information with tabular data on boundary changes to create a database of boundary changes with attribute data from the decennial census, employment reports, and other surveys, yielding a database permitting the study of historical change over the *longue durée* and across broad geographic space. Whereas historical geographers and historians had to restrict their investigations to local or regional case studies because the collection of data and manual cartography for larger study areas and greater historical depth were unfeasible, the GBHGIS and its counterparts in other countries vastly increase the historical territory that we can now explore.

3. Visualisation on the internet

One kind of exploration results from modifying the GBHGIS for public use on the internet. Entitled the *Vision of Britain through Time,"* this website offers "life-long learners" a great range of information about their locality. [4] The site design encourages the user to start by entering a place-name or postcode (the British version of a zip code), which yields statistics about the chosen district. Because the majority of the data have been converted to standardized units, the system can produce graphs of long-term trends within each district. Maps of attributes are also easily created. In addition, the user can explore information on specific places collected from historical gazetteers and contemporary travel accounts of the time ("travellers' tales." This rich and readily accessible resource has proved very popular among the general public. (Southall 2006)

As an academic resource, however, the *Vision of Britain* is of limited use. Firstly, the way that the data are structured is suited to studying an individual locality rather than regional or national patterns. Secondly, the system is highly data-driven, leading user to focus on the information itself rather than on research questions. Thirdly, because most of the statistical data have been re-mapped onto modern districts for easy comparisons over time, two problems arise. The modern districts are highly aggregated units (some 330 for England & Wales) so that detail preserved in smaller units like the parish is lost, masking rural-urban contrasts. Regrettably, the method of re-mapping and standardizing the data is not described so it is impossible at the moment to determine the accuracy of data. To permit scholarly use GBHGIS underlying the *Vision of Britain*, these issues can be resolved, as the following section shows.

4. Methodologies to integrate data over time

Ideally, when we explore change within a society we want to be able to handle time and space in an integrated way. Time is important because it allows the researcher to tell the story of how a phenomenon developed. If there is only one place however, be it a single village or a country single as an un-divided entity, there can only be one story. With a geo-referenced database one can explore multiple places to see how they developed along similar or different paths, a central scholarly concern (Massey 1999 and 2005). This examination of variation and difference is doubly beneficial. It avoids the pitfalls of research associated with logical positivism where the aim is to find a single model that can explain behaviour in all places under study. It also looks to uncovering the historical complexity inherent in temporal and spatial developments.

[Figure 5 – source and target parishes]

Within the GBHGIS the key to this is the ability to directly compare data over time at the most detailed spatial level possible. Unfortunately this cannot be done with the raw data because the myriad of boundary changes that affect parishes and districts over time make it almost impossible to compare one decade with the next, let alone to develop long-term time series. GIS offers a solution to this based on a set of techniques known as areal interpolation where the aim is to achieve “the transfer of data from one set (source units) to a second set (target units) of overlapping, non-hierarchical areal units” (Langford et al. 1991: p. 56). Figure 5 shows an example of the problem focussing on parishes in Wales. Parishes from 1911 have been selected as the target units and source population data from 1871 are to be interpolated onto these. The figure shows that the target parish of Pontypridd did not exist in 1871 and thus its 1871 population must be estimated from the populations of four source parishes: Llanwonno, Eglwysilan, Llanwitfardre and Llantrisant. Using a GIS *overlay* operation allows us to calculate what proportion of the area of each source parish intersects with Pontypridd. Using the assumption that the populations of the source units are evenly distributed across their area it then becomes possible to estimate the population of each source unit that should be allocated to each target unit. For example, in 1871 Eglwysilan source parish had a population of 8,200 and an area of 57.05km^2. The overlay operation reveals that 15.25km^2 of the source parish, lies in Pontypridd target zone, representing 26.7% of its area, the remainder lies in Eglwysilan target zone. This means that we estimate that 26.7 % of 8,200 people should be allocated to Pontypridd target zone giving 2,191, while the remainder is allocated to Eglwysilan. Using the same approach for the other three source parishes allows the 1871 population of Pontypridd target zone to be estimated. Conducting this for the whole country and for a large number of dates allows us to create national-level time series of data on a single set of target units. Although this involves significant amounts of computer processing time, this is not difficult to achieve.

The difficulty with this approach is clear: the assumption of an even population distribution is simply unrealistic. Among the ways of modifying this assumption, a technique known as the EM algorithm significantly improves the accuracy of the interpolated populations. The assumption here is that data from the target units can provide information on the intra-source zone distribution of population. Thus if data from a source zone is to be allocated to two target zones, one of which is urban and the other rural, a higher proportion of its population per unit area should be allocated to the urban target. The EM algorithm determines the relative proportions to use in making the allocations (Gregory & Ell 2005).

Another problem remains: how to determine the degree of error in the interpolated values. Estimating the likelihood of error in each data value can be done using techniques that identify unexpectedly large population changes that occur when boundaries are modified. Comparing the size of interpolated population changes with those that ignore boundary changes indicates the likely reliability of the data values (Gregory & Ell 2006). Painstaking interpolation, then, is well worth the effort: by eliminating or reducing the confounding effect of changing boundaries, the precision of the data and the reliability substantive results are considerably improved.

5. Tools for analysing change: Railways and population growth

To put these methods into action, this section takes up a concrete question: when, where, and to what extent did the expansion of railways in Wales affect population growth there from 1841—in the first decade of the railway age in Britain—to 1911 when the rail system arguably reached its geographic peak. In a rural area the coming of a railway could have both positive and negative effects. On the positive side, improved accessibility to high speed transport could stimulate the local economy by making it easier for producers to sell goods in distant markets and to profit from increasing production. On the negative side, the growing national rail system eventually increased external competition, undermining local production in favour of cheaper goods produced elsewhere. Although Wales lacks detailed information on the rural economy, parish-level population totals allow us to explore in detail the relationship between the rail transport and population change.

[Figure 6 – Growth of the rail network]

Railway lines, with their opening and closing dates, were digitised from a one inch-to-the-mile historical atlas of British railway development. (Cobb, 2006). Figure 6 shows how the network developed over the years 1851, 1871, 1891 and 1911. In 1851 lines had reached into the south of the principality, linking the major cities of Swansea, Cardiff and Newport with England. To the north, a major line ran along the coast to the port of Holyhead from which ferries carrying the Irish Mail departed and returned. By 1871 a massive expansion of rail lines had taken place, bringing rail transport to more and more localities and much of the whole principality. The rail network was especially dense in the southern Welsh Valleys where it served vast coal mines and related heavy industries. The maps of 1891 and 1911 show that the network did not expand much after 1871, growth being limited to connecting some existing lines and to reaching into rural areas along the south-western coast.

To compare the growth of the railways with population change, decennial census data for Welsh parishes from 1841 to 1911 were interpolated onto a single target geography, that of the 1911 parishes, of which there were 1,240. Parishes serve particularly well as the units for this type of analysis because they usually covered only a single village and its hinterland.

[Figure 7: Railways into parishes]

[Table 1: Spread of the railways]

Figure 7, the result of a GIS overlay operation, shows when and where certain parishes received rail transportation. Not surprisingly, the regions of greater economic importance in the north (slate and lead mining) and the south (coal mining and metallurgy) obtained rail service early on, while inland, agrarian regions joined the railway later, beginning in the late 1860s. The southern half of the west coast [when?] stands out as having very little rail transport. In Table 1 information from the overlay has been used to calculate the number and population of parishes that gained a railway in each decade. In 1851 railways reached only 11.8% (round to 12?) of the parishes. Not surprisingly, the parishes so served tended to be the most populous ones. All told, a third of the Welsh population lived in parishes with a railway.

By the end of the 1860s, after two decades of rapid expansion, rail lines existed in 45% of parishes which together comprised 62% of the total population of Wales. After the late 1860s and early 1870s growth slackened dramatically: over the next four decades only 107 parishes, with an additional 4.3% of the population, gained a railway, fewer than in any of the preceding four decades. By 1911, when expansion had all but stopped, the network covered just over half of the 1, 240 Welsh parishes and two-thirds of the total population. (After this date only a further three parishes gained lines for the first time.—put in a footnote?)

[Figure 8: Urban parishes]

To assess the impact of rail service in rural parishes we first have to define urban parishes (those with more than 10 persons per acre in 1911) and remove them from the analysis. Defined in this way, only 14 parishes in Wales fell into the urban category. All told, the fourteen contained 277,000 people, almost 15% of the Welsh population. These areas are shown in figure 8.

[Table 2: Pop growth]

Table 2 shows the relationship between the arrival of the railway and parish-level population growth. In the parishes where a railway first arrived, population growth was rapid the 1840s, a trend that continued in subsequent decades. In parishes that were newly served by a railway in the 1850s, the rates of population growth, already above the Welsh average before railways arrived, increased still further, before falling back had already [in later decades?] to pre-railway levels. Parishes that saw their first rail service in the 1860s and 1870s exhibited a pattern of growth but at lesser rates, reaching the national average or thereabouts and then declining. The 1880s experienced a different cycle of growth and decline. The high growth rates of the pre-rail decade fell in the 1880s when rail service first reached this group of parishes, and then turned upward in the 1890s. The relationship for the 1890s again conforms to the dominant pattern, with the arrival of rail associated with increased rates of growth that moved above the national average. In the 1900s the positive if temporary effect of railways on population change seems to have spent itself, for the few parishes that were newly served by rail differed little if at all from those where the "iron roads" never arrived: consistently low rates of growth. [no negative rates?] .

[Figure 9: Pop growth]

Figure 9 builds on this by showing the time series graphs of the parishes by the decade in which rail transport arrived. The extra detail shows some intriguing patterns. Parishes with rail service in the 1840s experienced the highest rates of growth while the rates fell markedly in the next decades. The 1850s and 1860s exhibit a similar pattern. suggested by table 2, population growth increases in the decade that the railway arrives but then drops back to near what it was before the railway arrived. The 1870s, 1880s and 1890s show a particularly intriguing trend in that in each case population growth increases in the decade before the railway arrives but falls away again. Only in the 1890s does it appear that rates after the railway had arrived were higher than rates before it arrived. Finally the 1900s and "never had a railway" parishes show patterns of growth that are consistently below average. [this seems to conform to the patterns described above. Can you boil it down to a few sentences, highlighting only the details that were not suggested in the previous analysis.]

In sum, rural places that were connected to the rail network early on experienced higher rates of population growth than other places would experience for the next half a century. To the extent that population growth indicates economic activity, then these early recipients of rail service benefitted more than other places both in the short and the long term. After the 1850s [1860s?], parishes that became connected to the growing regional and national rail network appear to have benefitted less and for a shorter period of time: after a sudden population increase the short-lived growth fell away. In Wales at least, the positive link between rail transport and population growth largely disappeared in the 1880s. This was a likely consequence of the agrarian crisis (1876-1894) when international competition in foodstuffs markedly intensified, and rising imports of cheaper grain from the United States and Canada, depressed agricultural prices generally and struck cereal farmers especially hard.

6. Conclusions: Towards a contribution to knowledge in the humanities

This discussion of Wales points to some of the strengths and limits of a study based on a National Historical GIS. On the positive side, a national historical GIS provides the data needed for detailed investigations of long-term historical change over extensive geographic space. It serves well as a means of integrating data from census returns, railway atlases, and many other sources, making it possible to pursue research questions that used to be beyond our grasp. With initial results in hand, complexities can be addressed by including additional data such as the locations and opening dates of railway stations, the volume of traffic on any given line, district-level information on employment, migration and agriculture, and digital models of the terrain—to name a few.

Like any approach, one based on historical GIS has its pitfalls and shortcomings. A notable pitfall is the limiting of research questions to fit the available data in GIS form. Among humanities scholars, this kind of reductionism is the target of much criticism and rightly so. A second criticism is that the findings are merely descriptive. They indentify what is where and when but tell us little or nothing about why. In other words, more is needed to explain the changing relationship between railway expansion and population change. We agree.

These and other criticisms call usefully for combining a GIS approach with other methods of geo-historical research, including those long favoured by many humanities scholars. Indeed, when it comes to explaining geographic patterns in the past, tried and true conventional methods of acquiring local knowledge are indispensable. In case at hand, marshalling local knowledge about Wales and its regions is an essential step toward an explanation of the patterns revealed by the GIS. The expansion of slate mines in the north, the development of coastal tourism, the growing demand for meat among miners and urban workers, the associated expansion of pasture in Pembrokeshire, and the increase in herds of store cattle raised in Central Wales and then sent south by rail—all these historic details are important parts of the story and explanation. [My sense is that this is a good ending without the following example]

To give a simple example, the analysis above is capable of identifying places that seemed to benefit from the railway, others that showed less benefit, and places that never received a railway. Benefit in this case is defined purely as population growth. This could complemented by an analysis in a local record offices or using local histories that perform detailed investigations of the impact of the railway specific places. The number of places that could be chosen is necessarily limited by time and by the records available. These case studies would tell us far more about the impact that the railway had on these specific places. The GIS in turn provides a framework for then deciding which other places are likely to have had similar experiences, and which did not.

Acknowledgements:

This paper benefited from support from the National Endowment for the Humanities under grant RZ-50577-06 and from the European Science Foundation's *Inventing Europe: Water*, *Road and Rail* Collaborative Research Project.

Boundaries for the 1850s parishes were derived from the work of R. Kain and R. Oliver (University of Exeter) converted into GIS form by N. Burton, J. Westwood and P. Carter, distributed by the UK Data Archive, Colchester, Essex as Study Number 4828. The

railway data from M.H. Cobb's *The Railways of Great Britain: A historical atlas*, where digitized at the University of Lleida by a project team led by Prof. J. Marti Henneberg.

For full details of staffing and funding of the Great Britain Historical GIS project see: http://www.gbhgis.org.

Shading schemes used on some of the maps in this paper are based on the Colorbrewer website (http://www.colorbrewer.org) produced by M. Harrower and C. Brewer.

References:

Berman M.L. (2005) "Boundaries or networks in historical GIS: Concepts of measuring space and administrative geography in Chinese history" *Historical Geography*, 33, pp. 118-133

Bol P. and Ge J. (2005) "China Historical GIS" *Historical Geography*, 33, pp. 150-152

Cobb M.H. (2006) *The Railways of Great Britain: A historical atlas. 2nd edition*. Ian Allen: Shepperton

De Moor M. and Wiedemann T. (2003) "Reconstructing Belgian territorial units and hierarchies: An example from Belgium" *History and Computing*, 13, pp. 71-97

Fitch C.A. and Ruggles S. (2003) "Building the National Historical Geographic Information System" *Historical Methods*, 36, pp. 41-51

Gregory I.N. (2002) "Time variant databases of changing historical administrative boundaries: a European comparison" *Transactions in GIS*, 6, pp. 161-178

Gregory I.N., Bennett C., Gilham V.L. and Southall H.R. (2002) "The Great Britain Historical GIS: From maps to changing human geography" *The Cartographic Journal*, 39, pp. 37-49

Gregory I.N. and Ell P.S. (2005) "Breaking the boundaries: Integrating 200 years of the Census using GIS" *Journal of the Royal Statistical Society, Series A*, 168, pp. 419-437

Gregory I.N. and Ell P.S. (2006) "Error sensitive historical GIS: identifying areal interpolation errors in time-series data" *International Journal of Geographical Information Science*, 20, pp. 135-152

Gregory I.N. and Ell P.S. (2007) *Historical GIS: Technologies, methodologies and scholarship.* Cambridge University Press: Cambridge.

Gregory I.N. and Healey R.G. (2007) "Historical GIS: Structuring, mapping and analysing geographies of the past" *Progress in Human Geography*, 31, pp. 638-653

Knowles A.K. (2005a) "Emerging trends in Historical GIS" *Historical Geography*, 33, pp. 7-13

Knowles A.K. (2005b, ed.) "Reports on National Historical GIS projects" *Historical Geography*, 33, pp. 293-314

Kunz A. and Boehler W. (2005) "HGIS Germany: An information system on German states and territories from 1820 to 1914" *Historical Geography*, 33, pp. 145-147

Langford M., Maguire D. and Unwin D.J. (1991). "The areal interpolation problem: Estimating population using remote sensing in a GIS framework." In Masser I. and Blakemore M. (eds.) *Handling Geographical Information: Methodology and potential applications*. Longman: Harlow, pp. 55-77

Massey D. (1999) "Space-time, 'science' and the relationship between physical geography and human geography" *Transactions of the Institute of British Geographers: New Series*, 24, pp. 261-276

Massey D. (2005) *For Space.* Sage: London

Southall H.R. (2006) "A Vision of Britain through Time: Making sense of 200 years of census reports" *Local Population Studies*, 76, pp. 76-84

[1] Gregory (2002) and Knowles (2005) provide general reviews of some of these systems.
[2] The China Historical GIS (Berman 2005; Bol & Ge 2005) provides an alternative approach based on the locations and spheres of influence of settlements. This allows them to cover a very large country for nearly two millenniums but, because it uses such a different approach, it will not be considered in detail here.
[3] Numbers of units are highly approximate as these change significantly with date
[4] See http://www.visionofbritain.org.uk