JOINED-UP THINKING ACROSS THE IRISH BORDER: MAKING THE DATA MORE COMPATIBLE

Rob Kitchin, Brendan Bartley, Justin Gleeson, Mick Cowman, Stewart Fotheringham and Chris Lloyd

Rob Kitchin

Brendan Bartley

Justin Gleeson

Levels of co-operation between government bodies, semi-state agencies, community groups and private industry in Northern Ireland and the Republic of Ireland are increasing rapidly with respect to issues such as coordinated infrastructure planning, health care, trade and enterprise, as well as peace and reconciliation. These initiatives require joined-up thinking and detailed rationales for their formation and funding, which in turn are dependent on evidence-based arguments. However, at present, providing consistent, coherent and reliable evidence on a cross-border basis for use at different scales is extremely difficult due to issues of data interoperability across the jurisdictions. In other words, data are often incompatible and therefore difficult to compare. Addressing issues of data interoperability is crucial to successful cross-border, evidence-based initiatives.

There is a reason why it is relatively rare to see all-island maps of Ireland other than those in a road atlas - they are very difficult to produce. And when they are produced, they usually have a small print warning about interpretation: something has had to be fudged to be able to create them. This is not to say that detailed information for both jurisdictions does not exist – it often does – or that there has not been a wealth of spatial analysis undertaken in the North or South – there has – but rather that data in the two jurisdictions

are largely incompatible as they are presently produced. As a consequence, undertaking cross-border or all-island spatial analysis faces a number of technical and other challenges. In this paper we outline these challenges and detail the work of the Cross Border Regional Research Observatory (CBRRO) in seeking to address them.

Poor interoperability

Interoperability in this context concerns the extent to which datasets that have been sourced separately can be used in conjunction with each other. If two sets of data cannot be used together because they do not share common attributes, then they are said to have very poor interoperability. Given that it is highly desirable that datasets can be combined so that sophisticated analyses can be undertaken, a series of different approaches have been adopted to try and ensure strong interoperability. These approaches include: (1) data agencies working closely together to ensure compatibility with regard to things such as data definitions and spatial units; (2) the establishment of various international data standards and conventions; (3) a drive towards national and transnational spatial data infrastructures that provide common frameworks and standards across borders and areas of concern (e.g. health, welfare, the economy); (4) the development of detailed metadatabases (data about data) that document what data are held by different agencies and their attributes; and (5) the development

of common data formats for recording and storing datasets so that they can be easily conjoined.

In the case of the Republic of Ireland and Northern Ireland, these initiatives have largely been confined to the level of the nation state and only recently has there been any real concern to improve cross-border interoperability. As a consequence, two separate approaches to data generation have developed, leading to poor interoperability on a number of levels. These problems exist in relation to nearly all types of data – including those relating to health, economy and enterprise, transport, environment, planning and development – but we confine our discussion here to a fundamental source of demographic and socio-economic data, the Census of Population. The census in Northern Ireland is administered by the Northern Ireland Statistics and Research Agency (NISRA) and in the Republic by the Central Statistics Office (CSO). While both censuses seek to record very similar types of data, they differ in a number of important respects.

1. The questions being asked

Although the censuses in Northern Ireland and Ireland share a common legacy and appear to be quite similar, they are prepared largely independently of each other. While many of the questions that are asked are either directly the same or very nearly the same in wording, a substantial number of questions examine the same issue

differently or there is no equivalent question. Of the 1161 SAPS (Small Area Population Statistics) variables outputted from the 2002 census in the Republic of Ireland, our analysis estimates that 32 percent of variables can be directly matched to the Northern Ireland census, 31 percent can be part-matched or reclassified so that they broadly match, and 37 percent have no equivalent. This means that over two thirds of all census variables published in the South have no *direct* equivalents in Northern Ireland without manipulation. For those issues where similar but different questions are asked, it is important to note that caution is necessary in comparing and interpreting the resultant answers across the two areas. This is because the question can be measuring highly related but subtly different phenomena or because the choices given to respondents do not match precisely.

Accordingly, a reclassification of answers might be needed in order to achieve a more meaningful correspondence.

2. Data units and categories

Similarly, the data being collected in the two jurisdictions might be recorded into different data units (e.g. euros instead of sterling), or into different data categories, or be outputted into varying classes. This effect is illustrated in **Table 1** which shows the different categories into which people are classified with respect to social class/grade in the Northern Ireland and the Republic of Ireland. While the categories are broadly similar, there are some significant differences that make straight comparison problematic. For example, semi-skilled and unskilled are separate classes in the Republic but are classified together in Northern Ireland.

Table 1: Comparing social class/social grade	
RoI	*NI*
Professional Workers *Managerial and technical* *Non-manual* *Skilled manual* *Semi-skilled* *Unskilled* *All other gainfully occupied grade workers*	*AB. Higher and intermediate managerial / administrative / professional* *C1. Supervisory, clerical, junior managerial / administrative / professional* *C2. Skilled manual workers* *D. Semi-skilled and unskilled manual workers* *E. On state benefit, unemployed, lowest*

There are many other examples. Occupations classed into seven different categories in the broad classification (SAPS) for the Republic of Ireland are all classed into the same category in the broad classification for Northern Ireland (Elementary Occupation). The reverse can also happen: for example, in Northern Ireland a judge and a refuse worker are classed as Professional Occupation and Elementary Occupation respectively, but in the Republic of Ireland both are classified in the SAPS data in an all-encompassing 'Managing, Administrative, Executive and Government Workers' class. In these cases full compatibility can only be achieved through the creation of common output classes for both parts of the island and the reclassification of data, but this takes time and careful thought to minimise any validity issues.

3. Spatial scales

Nearly all data have spatial attributes that allow them to be mapped onto the territory to which they refer. In other words they have an underlying 'output geography'. The output geographies for Northern Ireland and the Republic of Ireland differ quite significantly below the NUTS 3 level (which are counties in the Republic and five areas slightly bigger than counties in Northern Ireland). In the Republic census variables are coded at Regional, County and Electoral Division (ED) level (along with other more specialised units such as Gaeltacht areas). In the North they are coded at District, Ward, Super Output Areas, and Output Areas (OA) – along with other areas such as Health and Social Service Boards, Education and Library Boards, and Parliamentary Constituencies. As illustrated by **Table 2**, the characteristics of these areas are quite different, with wards having populations on average significantly larger than EDs, and OA populations significantly smaller.

Table 2: Comparing output areas between North and South

| Unit	Number	Av. Population	Av. Size (KM sq)
ED	3414	1062	20.4
OA	5022	337	2.8
Ward	582	2895	24

Because data in the two jurisdictions are reported at different spatial scales, a scalar modifiable area unit problem arises. In short, spatially referenced data can be aggregated into zones of varying sizes. The level of aggregation affects what patterns are revealed because the internal variances within a zone alter as more data are added. What this means is that the same data outputted at different spatial scales can show remarkably different patterns and statistical relationships with other data (*see Fotheringham and Wong 1991[1]*). As we have discussed, in the case of comparing data between Northern Ireland and the Republic of Ireland there is *no common output area*. As a result, the aggregation effects on internal variances are different between the two jurisdictions, making them statistically difficult to compare.

The visual effect of this is displayed in **Figure 1**. Here, the map on the left shows the population count for Cavan and Monaghan EDs and Fermanagh OAs, and the map on the right shows the population count for Cavan and Monaghan EDs and Fermanagh wards. In the first map, because EDs have significantly larger populations than OAs, the immediate inference one draws is that Fermanagh in Northern Ireland appears to have a uniformly lower population distribution than either Cavan or Monaghan in the Republic of Ireland. In the second case, wards, by and large, have a much greater population than EDs, and the opposite inference is drawn: that Fermanagh's population is uniformly higher than that of Cavan and Monaghan. In other words, the pattern that is displayed is not simply due to the population distribution, but is affected significantly by the spatial scale of output. Addressing this problem in the absence of a common spatial data unit is a complex process and is not easily resolved.

Figure 1: Comparing data outputted at different spatial scales: Fermanagh and Cavan/Monaghan
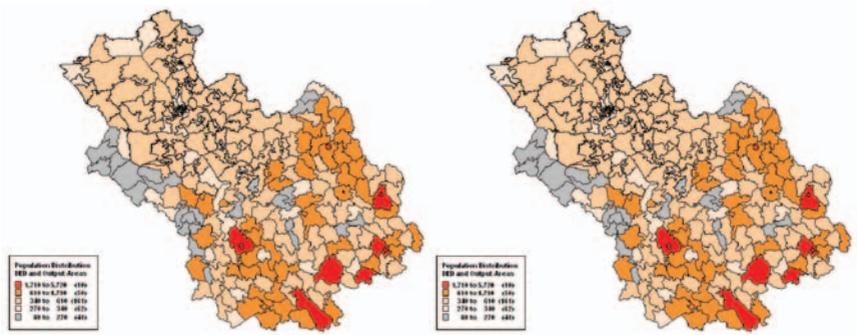
Beyond the census there are a number of significant issues with regards to *address matching*. While Northern Ireland has postcodes with a very fine spatial resolution (there are 56,114 postcodes averaging 20 addresses per postcode[2]), in the Republic of Ireland there is no equivalent, and because of the large proportion of non-unique addresses it is very difficult to pinpoint precisely an address even with the development of a new geo-referencing product, GeoDirectory[3].

4. Time series

There is another difference between the two maps in Figure 1. Because of the outbreak of Foot and Mouth disease, the census in the Republic of Ireland was delayed by a year. As a result, the data for Fermanagh refers to 2001 whereas the data for Cavan and Monaghan is from 2002. While a year seems a relatively short period, during this time the South was experiencing significant population gain. Indeed, a lot can happen in a short period, such as a large employer closing down and thereby altering local employment rates.

While the two censuses are aligned across a period of decades (1981, 1991, 2001/02, 2011), the census in the Republic also occurs on a five yearly cycle (1986, 1996, 2006, 2016). As a consequence, while the Republic of Ireland will have data relating to 2006, Northern Ireland only has data for 2001. While a year's difference might be manageable when making cross-border

comparisons, a five year gap is a significant time period, and comparing data collected in 2006 in the South with that collected in 2001 in the North would be highly problematic. Related to this is the problem of drawing on 2001 census data as time progresses. For example, by 2010 Northern Ireland will be, in many ways, quite different to how it was in 2001, and yet this will be the last full census from which to draw socio-economic data (although NISRA will have collected other related data in intervening years).

5. Data continuity

In order to be able to easily compare data across time periods, it is important that both the data generated and the areas into which they are outputted have continuity. In both Northern Ireland and the Republic of Ireland there are a number of continuity issues that need to be addressed. For example, in the Republic, for each census year between 1981 and 2002, the number of SAPS variables released has varied (from a low of 774 in 1981 to a high of 1750 in 1991) as some questions have been added to the census form and others discontinued. Clearly in these circumstances it is impossible to compare new or discontinued variables over time because no comparable data exist.

In Northern Ireland there have been significant changes to the spatial units for which data are outputted. For example, ward boundaries were altered

in both 1984 (566 wards) and 1992 (582 wards). As a result, while many wards remained the same between the 1991 and 2001 censuses, others had their boundaries altered or were merged with other wards, and some renamed[4]. It is a difficult task directly to compare data outputted into different spatial units because of discrepancies in the underlying geography. The boundaries were altered again in 2002, meaning that the 2011 census will have a different ward geography to previous censuses. Changes also occur at other scales, such as at district council level (e.g. in 1992 the ward of Rathfriland was transferred from Newry and Mourne Local Government District to Banbridge Local Government District), and at postcode level where boundaries are subject to continuous change due to new addresses and alterations to the built environment.

6. Context

In some cases census data differ, and the interpretation given to data needs to differ, because of the context in which the data are collected. Different policies and economies operate across the two jurisdictions, meaning that the issues the data purport to measure are shaped in different ways. For example, take home or disposable income varies as a function of the tax regime, not simply the level of earnings, and the number of people living in local authority housing is dependent on housing policy and public sector provision. In these cases a straight interpretation of the data that fails to take account of policy or regime will lead to erroneous conclusions.

7. Metadata and data clarity

In general, census data provided through NISRA and CSO has metadata – data about existing data, usually concerning its availability, organisation and use – that makes them relatively easy to understand and work with. Such summary information is not always available for other kinds of data. There are two general metadata issues. First, metadata on what data exists within both jurisdictions is patchy, with no one resource that lists all the sources of national and regional coverage data. Second, metadata about specific data sources is patchy about the composition (foci, variables and coverage) of potential data or issues relating to cost, licensing and usage. In some cases, data users have to discover this metadata for themselves, tracking down specific sources of information or by 'playing' with the datasets to find out what it comprises. While some census data are reported as raw data counts that can then be analysed by users as desired, it is not always clear what the data consist of in other cases. This is especially the case if data are not raw counts or percentages but are the outputs of a statistical formula. Without knowledge of how the data were derived it is difficult to interpret what they show.

8. Availability and sourcing

Adding to these concerns, and drawing from our own experience, it is clear that

there are substantial issues with regard to the availability of data and the sourcing of data in general. While these problems are limited with respect to census data, which are freely available for both jurisdictions, they clearly operate with regard to other data. Many data, especially relating to enterprise and innovation, are either not available in one jurisdiction or the other, or are only available at certain scales (e.g. at the national scale but not at the regional, county or sub-county scale). This clearly limits the analysis that can be undertaken and provides only a very weak and generalised view when seeking to make comparisons.

In addition, we have encountered problems trying to source data we knew existed because they are not openly published or are not published at the scales required. Further, in some cases it has not always been clear who is responsible for compiling and publishing data, making it difficult to track down. Related issues concerning data availability are cost, licensing arrangements and data formats. Data in both the UK and the Republic of Ireland tend to be relatively expensive and there are a number of restrictions on how the data can be used.

Addressing interoperability: The CBRRO

The issues discussed above are well known to those working within data agencies and they are certainly of concern to them. Several initiatives are underway to address the problems

noted, including the Irish Spatial Data Infrastructure (ISDI) committee and working groups in the Republic of Ireland; and the Mosaic initiative (geographic information strategy) in Northern Ireland, including Geohub (a central resource for spatial data for Northern Ireland). There are also four cross-border initiatives: (a) the Spatial Indicators project that provides cross-border land cover and use modelling using the Moland model; (b) the North-West Data Capture project that is seeking to provide a common cross-border dataset for planning in the Derry-Donegal region; (c) the Mapping INTERREG project, led by the International Centre for Local and Regional Development (ICLRD), that is seeking to spatially reference and map the funded projects detailed in the Centre for Cross Border Studies' Border Ireland database; (d) and the Cross-Border Regional Research Observatory (CBRRO), which we outline in more detail below. Despite these projects, it is fair to say we are at the start of a very long process in terms of improving evidence-informed decision-making and all are confronted by the problems detailed above.

The Cross-Border Regional Research Observatory (CBRRO), based in the Cross Border Centre for Community Development in Dundalk Institute of Technology, was established in February 2006, and funded as a pilot project by the Special EU Programmes Body for an initial six-month period. The initial aim of CBRRO was to work towards providing precise and reliable regional

29

intelligence that would enable a better understanding of the dynamics of the cross-border region, and aid the formulation of strategic policy development and cross-border cooperation. Good cross-border, regional intelligence is dependent on solid evidence. And as we have seen, there are a number of outstanding issues concerning the interoperability of cross-border data. To help address these issues throughout 2006, the CBRRO undertook six key tasks with a view to establishing a full-scale Regional Research Observatory that would operate for several years to come, working with data agencies and those that use their data.

First, a full review was undertaken of what datasets are compiled in Northern Ireland and the Republic of Ireland, the nature and format of the data, along with their source, cost, license arrangements, scale, and the extent of interoperability with comparable, cross-border data. In particular, a detailed examination was undertaken of the censuses in both jurisdictions to determine the level of compatibility between variables. Second, we established a baseline of what spatial analysis is presently undertaken within or between both jurisdictions, what kind of analysis is being performed, by whom and for what purpose. Third, we examined the work of other regional research observatories operating elsewhere, including their structures and functions, the research they undertake, their outputs, how they are funded, and so on.

Fourth, we undertook an additional study compiling a socio-economic profile of the Border counties, Northern Ireland and Western Scotland that enabled us to work with and examine data from a variety of sources beyond the census. Fifth, we have started to develop an initial set of tools designed to improve data interoperability and help data users understand and maximise their use of cross-border data. And finally we scoped out what a long-term CBRRO would look like and what it would seek to achieve.

It is most useful to concentrate here on the fifth of these points and some of the tools we have started to develop to help people access, make interoperable and analyse existing data sets. The CBRRO has developed a web-based interface to cross-border census data and associated metadata that consists of four key modules:
- a metadata portal
- a priority indicators module
- a mapping module
- a geographic profiling tool

The *metadata* portal is split into three components, each detailing information for both Northern Ireland and the Republic of Ireland: a directory that provides contact details and links to key organisations; a publications and reports repository providing direct access to key documents; and a spatial data inventory that consists of a metadatabase of all spatial-referenced data. The *priority indicators* module consists of access to a series of pre-prepared, interoperable

Figure 2: A sample of all-island, priority indicator maps showing (from top left to bottom right) – proportion of population over 65, proportion of people classed as Catholics, proportion of people with higher education, and proportion of people living in local authority housing
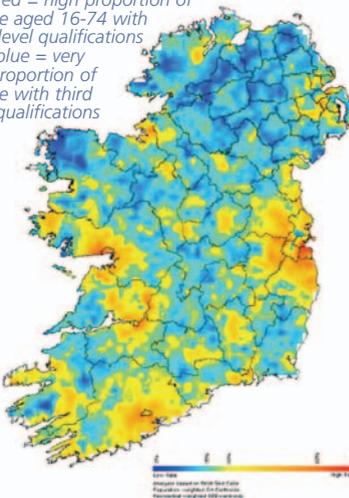

dark red = higher proportion of population over 65
*dark blue = very low proportion of population over 65*
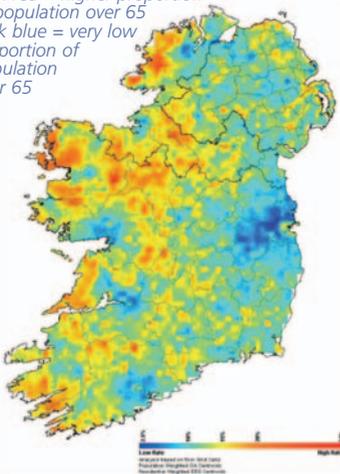
dark red = higher proportion of Catholics
dark blue = very low proportion of Catholics
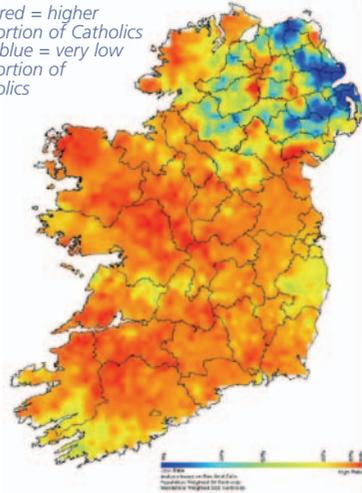
dark red = high proportion of people aged 16-74 with third level qualifications
dark blue = very low proportion of people with third level qualifications

dark red = higher proportion of people in local authority housing
dark blue = low proportion of people in local authority housing
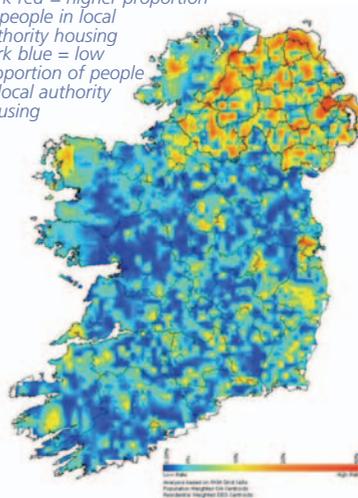
maps of key socio-economic variables (**see Figure 2**). The *mapping module* provides users with the flexibility to create their own maps of data pre-prepared so as to be interoperable. It uses off-the-shelf software, Instant Atlas, and users can alter the variables displayed, the output areas, filter data, and query the map. The *geographic profiling* tool enables relatively sophisticated querying of pre-prepared data by allowing users to define an area and to automatically generate data summaries for that area, even if the area straddles the border.

To tackle problems of EDs in the Republic of Ireland being intersected by the selected or defined area, a proportioning formula has been developed that uses the address database Geodirectory to calculate what proportion of the population within an ED live within the defined area and to weight the data accordingly. Within Northern Ireland, OAs are used, and because they are significantly smaller than EDs there is no need to use a proportioning algorithm. The data summaries produced consist of a standard text template, an excel spreadsheet with the results for the defined area plus other useful comparator regions (e.g. Northern Ireland, Republic of Ireland, counties and districts), and a set of charts and graphs.

It is hoped, subject to licensing arrangements, that these tools can be hosted on a live CBRRO website. In addition, the intention is that the CBRRO will be populated by more available data, making time-series analysis, along with analysis at scales from the local to the EU levels, possible. It is also hoped that other tools and outputs will be developed, including the ongoing tracking of key indicator variables, as well as the publication of regular regional intelligence reports and a newsletter.

Conclusion

Cross-border collaboration across a range of issues has increased substantially in the last few years. In turn, such collaboration has been accompanied by a desire that initiatives be underpinned by sound, supportive evidence that justifies investment and can reveal the benefits yielded by constructive collaboration. At present, however, it can be extremely difficult to provide such evidence because data are often poorly interoperable in a number of important aspects.

In this short paper we have tried to highlight the various ways in which data for Northern Ireland and the Republic of Ireland often lack interoperability, and why such difficulties are important to address. While the CBRRO and other related projects are starting to tackle these issues, it is fair to say that substantial long-term research and development is needed in order to achieve progress. Such progress is likely to be attained through the work of the CBRRO and related projects on the following five fronts, namely: (1)

preparing interoperable data across several domains; (2) addressing fundamental technical issues such as modifiable areal unit problems; (3) providing new, more sophisticated tools of analysis that work on an all-island or cross-border basis; (4) helping to educate data users; and (5) promoting inter-jurisdictional data analysis that will provide the kinds of evidence that will serve policy makers well.

Rob Kitchin is Director of the National Institute for Regional and Spatial Analysis (NIRSA) and Associate Professor in the Department of Geography at the National University of Ireland, Maynooth. He is also a board member of the International Centre for Local and Regional Development (ICLRD).

Brendan Bartley is Acting Director of NIRSA and a Lecturer in urban planning and geography at NUI Maynooth. He is also ICLRD programme director at NUI Maynooth.

Justin Gleeson is Project Coordinator at the Regional Research Observatory, NIRSA, NUI Maynooth.

Mick Cowman is Director of the Cross Border Centre for Community Development at Dundalk Institute of Technology

Stewart Fotheringham is a Science Foundation Ireland Research Professor and Director of the National Centre for Geocomputation, NUI Maynooth.

Chris Lloyd is a Lecturer in the School of Geography, Archaeology and Palaeoecology, Queen's University Belfast

ENDNOTES

1. Fotheringham, A. S., and Wong, D.W.S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning* A 23: 1025-44.
2. http://www.ninis.nisra.gov.uk/mapxtreme/linkeddocs/GeoreferencePaper_Dec2005.pdf
3. http://www.geodirectory.ie/index.html
4. http://www.nisranew.nisra.gov.uk/Census/metadata/Geography.html