Using Administrative Data to Identify Individual Risk Factors for Unemployment in Economic Downturn

Rhodes, Mary Lee
Trinity College Dublin, School of Business
College Green
Dublin 2, Ireland
E-mail: rhodesml@tcd.ie

Ferguson, David
SAS Institute, Business Advisory
50 Northumberland Road
Ballsbridge, Dublin 4, Ireland
E-mail:David.Ferguson@irl.sas.com

1) Introduction:

In this paper we present the results of a research exercise in linking records residing in multiple agencies / departments in the Irish government and applying statistical methods to the resulting dataset to demonstrate some of the practical advantages and challenges in exploiting existing administrative data. This analysis was undertaken by the Analytics Institute in Ireland working in collaboration with researchers in the Central Statistics Office in Ireland and Trinity College Dublin. While the specific data used are related to unemployment, the methods used and the challenges and opportunities identified are more generally applicable to a range of policy domains and public agencies in Ireland. In fact, the findings related to individual risk factors relating to unemployment in an economic downturn are preliminary only and would need further study before they could be relied upon to inform policy or programme implementation. The objectives of the research presented here are to:

1) demonstrate the value of available administrative sources, particularly when multiple agency sources are combined;
2) demonstrate the use of different analytic techniques for exploring data to address important social and economic issues;
3) demonstrate the value of different presentational techniques to 'tell the story' in a way that people can relate to and understand[1];
4) examine the challenges and opportunities in exploiting administrative datasets, and develop guidelines for future use.

In the next section (section two) we highlight features of the Irish socio-economic context that give rise to a number of important policy questions that inform the research exercise undertaken. In section three, a

[1] This objective was addressed in the presentation of the results at the ISI Congress 21-24 August 2011

general framework for data mining is described that was used to develop the specific steps taken to produce the dataset needed from existing administrative sources. The three main sources used are the Central *Statistics Office* Quarterly National Household Survey (QNHS), the *Department of Social Protection* Client Record System (CRS) and the *Department of Revenue* P35 Employment database (P35). In addition, descriptive data from the *Central Statistics Office* Business Register (BR) was incorporated into the final dataset. Section four describes the main analytic techniques applied to discover likely individual risk factors for unemployment in a downturn along with the findings and also provides examples of how the results may be presented. We conclude in section five with a reflection on the implications of the exercise undertaken for public policy and management practice and some proposed guidelines for future exploitation of existing administrative data in the Irish public sector.

2) Research context and questions to be explored:

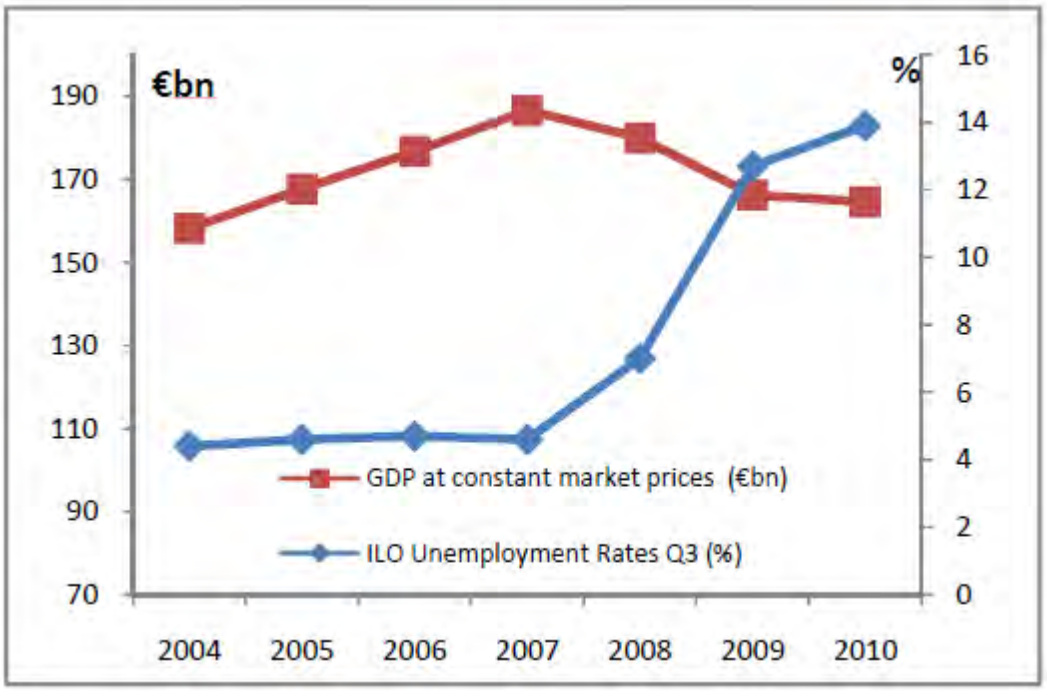
There are two main issues in public management and policy making in Ireland that provide the driving force for the research exercise described in this paper. The first is the need to ‗do more with less' – a phrase that appears in numerous policy documents and analyses of what is wrong and what needs to be done urgently if the Irish Public Service is to respond effectively to the fiscal crisis that has dominated the past three years of public policy. There are numerous reports that express this in various ways with three relatively recent ones being *Transforming Public Services* (2008) on the programme for public sector reform in Ireland, the *'An Bord Snip'* report[2] (2009) on public expenditure cuts, and *Fit for Purpose?* (2011) on the challenges and priorities for the Irish Public Sector. One way of increasing efficiency in the public sector would be to use existing administrative data more effectively, thereby decreasing expenditure on consulting and bespoke research projects. Exploiting the existing data through data linkage across departments and agencies and applying advanced analytic and statistical techniques could also result in more targeted programmes – reducing the incidence of ineffective or overly broad initiatives.

The second driver concerns the specific policy domain chosen for this exercise. As mentioned above, Ireland is in the midst of an fiscal and economic crisis that has resulted in a rapid deterioration of output and a related spiraling upward of the unemployment rate (see figure 1). Macro-economic drivers of employment are well documented and a great deal of research has been done to understand the causes of long-term unemployment (see Ljungqvist & Sargent 2008 for a recent discussion). Furthermore, there are man y studies of the demographic and behavioural factors contributing to the risk of being unemploye d (for references in the British context see Payne & Payne 2000). The financial and social suffering of people who are unemployed is well understood and there are numerous supports in place to mitigate this suffering as well as to help people back to work. In the context of the rapid and significant deterioration in the job market and the lack of government funding available, it would be of value to understand where to focus scarce resources to best effect in helping prevent people becoming unemployed in the first place, to

[2] This is the colloquial name for the *Report of the Special Group on Public Service Numbers and Expenditure Programmes (2009), Volume 1, Dublin: Stationery Office*

implement retraining and/or job-search programmes and to put in place appropriate supports. Hence, the policy domain for the analysis exercise is on unemployment and, specifically on the risk factors that indicate likelihood for becoming unemployed in a given period after the economic down turn.

Figure 1 (Source: CSO)



The specific research question that we used to focus our efforts was: "What are the demographic or behavioural risk factors that significantly increased an individual Irish worker's likelihood of becoming unemployed within 1 year, 2 years or 3 years following the 2007 economic collapse?" Note that this question evolved over the course of the project as we engaged in more depth with the data and the analytic potential of the models that were developed.

3) Creating the dataset from existing administrative data:

The process of acquiring and preparing the data required involved two major steps: 1) record linking and anonymisation of individual data from different administrative sources, and 2) creating a dataset in the form required for the analytic processes. **The first step** was largely carried out by analysts in the CSO and data matching across different sources was done in line with the Statistics Act, 1993[3]. For the anonymisation process, an individual's Personal Public Service Number (PPSN) was removed from the files, and replaced with an internal CSO identifier. Names and addresses were removed, and dates of birth were truncated to month of birth. All data must be anonymised before it is made available to 'authorised users' and to

[3] For the CSO Data Matching Protocol, see http://www.cso.ie/aboutus/data_protocol/data_protocol.htm

become an authorised user, an individual must be being sworn as Officer of Statistics under the Statistics Act.

With respect to record linking across the specific files used, the Quarterly National Housing Survey (QNHS) only began collecting the PPSN during 2007, and only as an optional field. To increase the coverage, the CSO performed a lookup of the QNHS participants on the Department of Social Protection Client Record System (CRS). The matches were based on date of birth, surname, and first name.

Unemployment claims were extracted from the CRS (‗Live Register') as well. Each claim was recorded once in the year it began. The file contained all claims made for Unemployment Benefit or Allowance, those signing for credits only, and those participating in back to work schemes.

The Revenue P35 data contained one record per person for each employment they held during the year. It contained details of total pay and PRSI amounts, and the number of weeks worked in each employment. It was linked to the CSO Business Register database to identify the NACE sector of the employers, and to the CRS to add demographic data on the individuals.

The second step in preparing the data for analysis was performed by one of the authors and involved creating a single ‗flat file' dataset from the anonymised files received from the CSO to facilitate the analysis undertaken. Data mining and analytical techniques work best when the data in a flat and non-normalised state. This means that each observation in the target data set contains the main outcome variable (in this case an indicator for employed / unemployed) and all the variables relating the individual record from the various data sources i.e.

Age Band	Gender	Region	Unemployed
20 – 30	Male	West	1
30 – 40	Female	East	0
Etc	Etc	Etc	etc

The research question being asked defined the target outcome variable and the various factors relating to an individual who was employed/unemployed were extracted from the different administrative data sources. The main linking identifier is the anonymised PPSN number, which we will refer to as 'CSOPPSN' going forward.

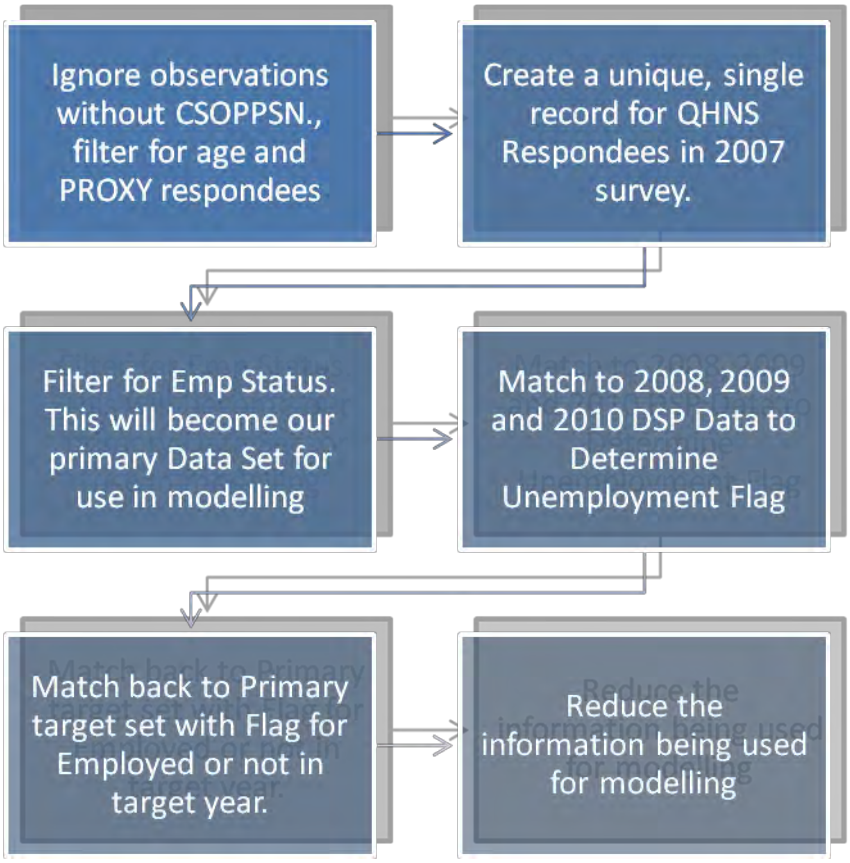
To address the research question and to ensure a relatively clean and comprehensive file, data was selected from the various sources using the following rules:

1. All records on the QNHS had to be identified with a CSOPPSN number for use in linking and matching to P35 and importantly could be matched to the Live Register data

2. Individuals ' primary economic status (PES) from the QNHS was 'employed' in 2007
3. Excluded individuals who were not of working age (<14)
4. It was possible to tell if the individual did or did not become unemployed in 2008, 2009 or 2010

In addition to the above, several steps were taken to create a datafile that would allow the particular data mining processes chosen to be run efficiently. A process flow of the data preparation phase (post CSO processing) is provided in Figure 2 and a description of each step is provided in Appendix 1.

Figure 2: Process Flow of data preparation post CSO extraction

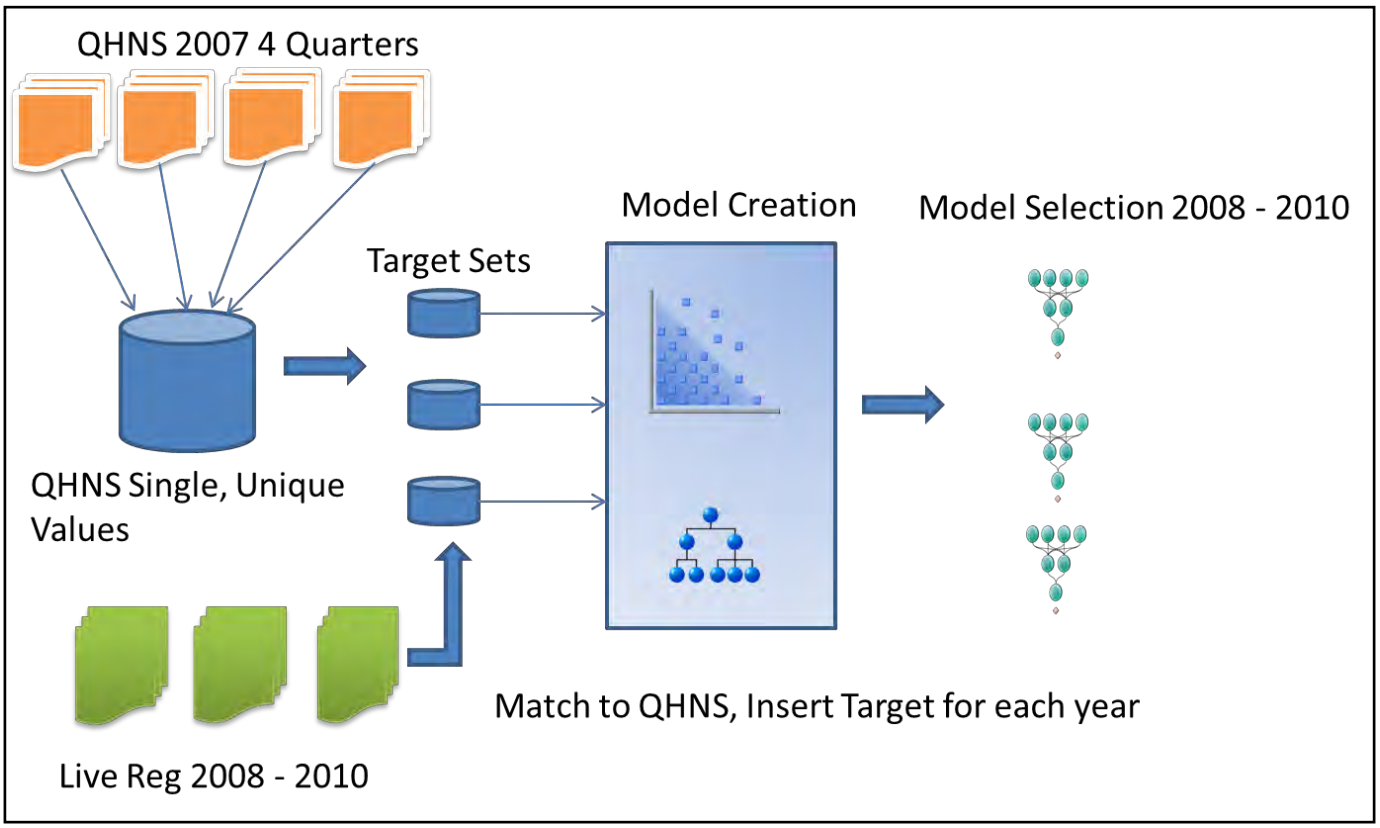


At the end of this process, we were left with a data set of approximately 44,000 records of individuals, of a total of over 196,000 records for the 2007 QNHS survey data sets, who were employed in 2007, who met the criteria (age / valid CSOPPSN) and who may or may not have lost their jobs over the following 3 years. Note that the specific proportion of individuals who became unemployed in each year was 11.6% in 2008, 17.1% in 2009, 13.8% in 2010. The aggregate of these percentages exceeds the total increase in unemployment (from 4% to 14% over the period) most likely due to individuals becoming reemployed after a period of unemployment in the three years of the study.

4) Analytic approach and data mining:

Data mining is no different to any business or technical process in that it works more effectively when a structured approach is taken to achieving the desired outcome. However, like most research processes, data mining is not a linear process. It is iterative and adaptive to results produced in previous iterations (feedback) and generally interactive between and among the analyst(s) and the analytic tools used. Figure 3 below represents the main tasks undertaken in producing the analysis. On the left hand side of the diagram we can see a representation of the processes involved in creating the data set followed by an iterative process of apply different modeling techniques (described below) to explore different models that could address the problem posed. The last iterative process was selecting the model (in the case the set of predictive factors for unemployment) that provides the best results (in this case the highest probability of differentiating between those individuals who are likely to become unemployed and those who aren't).

Figure 3: The full set of processes involved in the data mining research project



In this section we focus on the middle and right-hand sides of this diagram to describe the processes of model creation and model selection to answer the question, "What are the demographic or behavioural risk factors that significantly increased an individual Irish worker's likelihood of becoming unemployed within 1 year, 2 years or 3 years following the 2007 economic collapse?"

4.1 Model Creation Processes:

The process of creating an analytical model is very much determined by the business question being asked. There are a variety of data mining techniques that have their own uses and relevance to a particular situation. Given the research question we had developed, the first decision we made was to **classify an outcome,** i.e. whether the QHNS respondents who were employed in 2007 were unemployed or not in the years 2008, 2009, 2010. This step was taken during the creation of the dataset as described in the previous section.

The next decision was to use **supervised learning techniques** for exploring different models as opposed to unsupervised learning techniques. Supervised learning techniques are those typically used in classification or prediction type analysis. They use the fact that there is a known outcome and will _learn' from that outcome about the factors / predictors and the relationships between them[4]. Unsupervised techniques, which are used when there is no outcome to predict were not considered as we wanted to focus on the outcomes of employment or unemployment in our analysis.

To support the presentation of results and the creation of a _story' to support any hypotheses that we develop, we took the third key decision to, i.e. to utilise techniques that will provide a clear breakdown of the factors and their relationship with our target outcome. In the end we used **decision trees and regressions** to analyse the dataset and develop a list of predictive factors and hypotheses in relation to the rolling effect of an economic crisis on different subsets of the working population.

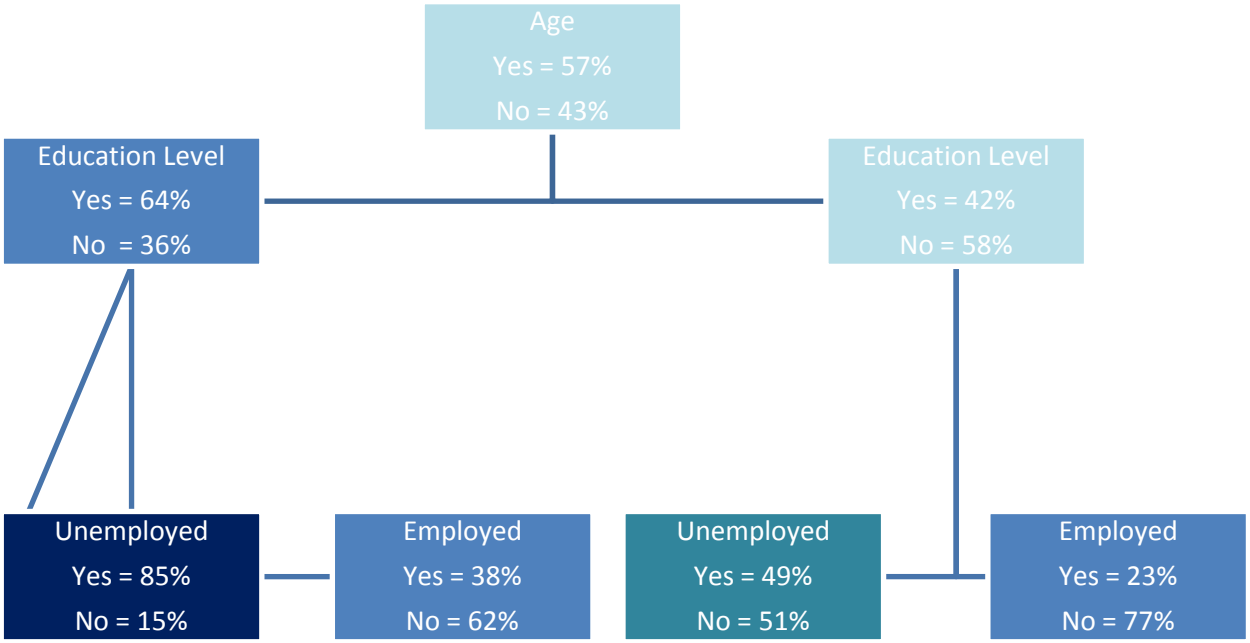
To create a comprehensive set of models that describe the factors influencing unemployment for the years 2008 – 2010 we decided to create a model for each of the years separately. The reason for this decision was to look at changing factors over that time period, i.e. had the factors changed for people who were employed made unemployed in each of 2008, 2009, 2010. In addition to understanding which factors were more influential in each year, this also allowed us to create hypotheses about the _rolling' effect of an economic crisis and its impact on different segments of the population.

Decision Tree Models

A decision tree is a modeling technique for composing a set of decisions that explain the relationship between a number of attributes and the target decision (in our case the classification of Unemployed). The decision tree is composed of a branch and node structure, where a node represents a decision to be made on the attributes available, according to specific criteria (either **statistical** or **heuristic**) and the branches represent the decision outcomes. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. Figure 4 is an example of the sort of result that a decision-tree model might produce from the data.

Figure 4: A hypothetical decision-tree model output

[4] See Shmeuli, Patel and Bruce (2010) Data Mining For Business Intelligence (2nd Ed) for full explanation of supervised and unsupervised techniques.



Some of the key criteria in relation to how we went about creating a decision tree from the data are:

Splitting Rule (i.e. the rule basis used to determine how to split on a particular variable (such as Age above): p-value of Pearson Chi Square Statistic for the target of unemployed vs. the attribute being tested. Other criteria available and relevant for nominal variable classification are reduction of variance measures such as Entropy.

Decision Criteria: Minimise Misclassification rates as opposed to other criteria such increased classification ‗lift' per decile of the target population.

Note that one of the strengths of decision tree models for exploring data is that they do not suffer from instability from class variables or non-normal data as much as other techniques. However, this positive feature of being more forgiving of dodgy data is offset by generating less precise relationships and models that capture the full set of relationships between independent variables and the target outcome. This trade-off is also present (in the opposite direction) for regressions.

Regression Models

Regression analysis is a statistical method to determine what factors have a strong association with a target variable. Regression analysis creates a prediction formula that weights each factor in that formula. They can be used to heuristically evaluate a target using approaches such as stepwise, forward selection and backward selection. Note that for this project as our dependent variable is a binary variable (regressions were initially designed to predict continuous variables), we used a *Logistic Regression*. In logistic regression, the expected value of the target is transformed by a link function to restrict its value to the unit interval (between 0 and 1). In this way, model predictions can be viewed as primary outcome probabilities. A

linear combination of the inputs generates a logit score, the log of the odds of primary outcome, in contrast to the linear regression's direct prediction of the target. Some important parameters selected for the Regression Analysis:

1. Logistic Regression
2. Stepwise Logistic Regression
    a. Significance level of .05 for entry of effect
    b. Significance level of .05 for staying in the model as an effect

Regression algorithms are more sensitive to high dimensionality of the data and/or missing data than are decision trees. Due to the data set preparation step, there was very little missing data, and the only issue was the high dimensionality one – particularly in relation to the NACE level 2 Detail codes. Therefore we used two regression algorithms. The first we linked directly to the data so it was receiving the full set of information from our target set. The second we ran following a Decision Tree analysis using the outputs / variables from the decision tree. This second method helped to reduce the number of variables and redundant predictors (according to the Decision Trees outputs).

As noted above, we used a commercial data mining product called SAS Enterprise Miner. This tool was useful in that it allowed us to create a process flow of the different algorithms, data partitioning, reporting and model assessment using a GUI interface. In addition, it automated the processes of establishing and testing the models through the creation and application of: 1) the *Learning* (training) dataset, 2) the *Test* dataset and 3) the *Validation* dataset. The learning or 'training' dataset is used to develop the set of relevant variables (algorithm) as a first pass through the data. The test dataset is used to test the algorithm created from the training set and the validation dataset is used to validate the model that is created by comparing the different algorithms developed in the learning and test datasets. SAS Enterprise Miner 6.1 allows us to automatically split the data into different proportions. We chose a 40:30:30 proportion split, which provided each set of data with a good representative number of observations of the target variable per set.

4.2 Model selection and assessment processes

There are many different methodologies for assessment between models. Using SAS Enterprise Miner allowed us to automate the assessment of the models using a parameterised assessment step in its processing. There are a number of different options available to make an assessment between models, given our task was classification, using Misclassification Rate and ROC (see below) were most appropriate. For other options available in SAS Enterprise Miner see Appendix 3.

There are a number of ways of assessing the efficacy and robustness of the model for classification type tasks. The criteria we selected for assessing the efficacy and robustness of the models produced were:

1. Misclassification rate (minimize the proportion of false positives and false negatives);
2. Receiver Operating Characteristic (ROC) (maximize the ratio of true positive predictions to false positives);

Both these are closely linked as they are related to misclassification as an assessment criteria to select between models.

3. Average Squared Error (minimize this for each regression tried);
4. Differences between validation and test scoring outputs

To avoid model overfitting or underfitting (i.e. a model that is too specific and will not work on new information or a model that is too general and is no better than selecting events at random) we constantly monitored the differences in outputs (such as misclassification scores, Average Square Errors etc) between the validation subset of data and training subset of data. This is a criteria we used while testing different models / techniques to avoid the models overfitting or underfitting.

Note that the misclassification rate was our key model assessment criteria, as this gives us an indication as to how successful the model will be at differentiating between those individuals that are likely to experience a period of unemployment in the year selected and those that are not. This was a key issue in the study reported in Payne & Payne (2000) and is an important consideration in the application of any findings to policy development and/or practice.

5) Findings and implications for policy / practice:

In this section we present the findings from the model selection process as factors that may be used for predicting unemployment of individuals following an economic crisis. Note that we do not suggest that these be taken at face value, but rather that they provide a strong basis for further testing, as well as data that may be of use to policy makers and practitioners as part of a rich tapestry of information on which to develop appropriate interventions. We also reflect on the opportunities arising from this type of exercise for the public sector and some of the challenges that we faced in pursuing same. In the presentation of findings we first discuss each year separately and then propose a hypothetical ‗story' to explain the findings across years. Note that in each year we have identified the top 3 predictors for becoming unemployed.

Predictors of being unemployed in 2008

The most significant predictor was *occupation* (NACE2), with 5 occupations having a high risk of

unemployment: MANUFACTURING[5], FABRICATION[6] CONSTRUCTION[7], HOTELS[8], and MINING[9]. The second most significant predictor was *age*, with those in the age grouping of 19-24 having the highest probability for becoming unemployed. The third most significant predictor was *income level*, with individuals earning between €391 and €720 per week being at risk of unemployment in the year immediately following the crash. While it is well know that construction and hotels are risky (and cyclical) businesses, it was somewhat surprising that manufacturing and fabrication would also be so quick to shed employees. Mining was a surprise – but may have been driven by the closure of Tara Mines around that time and the loss of around 800 jobs. Youth is also a well-known 'risk' for unemployment, but the income level of between €20-37K per annum was an unexpected finding as one might expect that lower paid workers are at most risk immediately following a crash. Given the type of occupations that were identified as most at risk, however, this could simply be a correlated risk factor with the primary one of occupation.

Predictors of being unemployed in 2009

Again, *occupation* in 2007 is the most significant predictor of becoming unemployed in 2009, but in this year HOTELS disappears from the list of high risk jobs and is replaced in the top five by REALESTATE[10]. This suggests that there was a rolling effect of the crisis across all jobs relating to property – beginning with construction (and possibly the building supply industry in so far as it is part of Manufacturing and/or Fabrication) and moving into real estate. This is not unexpected given that a major contributor to the economic crisis was the property bubble in Ireland. However, it must be noted that the other big contributor to the crisis – reckless lending, primarily to the construction / development sector – did not appear to have a similar impact in terms of unemployment for those involved in banking. The resilience of employment in the banking / finance industry in Ireland is worthy of further research. The disappearance of HOTELS as a high risk occupation suggests that this sector is able to shed the necessary level of jobs quickly – or possibly that the industry recovered more quickly.

Even more interesting than the shift in occupational risk, is the change in the second and third highest risk factors for unemployment in this year. As the second highest risk factor for unemployment in 2009, age is replaced by *education*, with those individuals having achieved only up to and including second level qualification at significant risk. While this is not surprising at one level – since it is well known that education is a significant factor in employment level and remuneration – the fact that it appeared only in the second year after the crisis is interesting and worthy of further study. The third highest risk factor had to do with *family cycle* – whereby families with young children and 'non-family' households were at a significant risk of unemployment in this year. This also suggests a rolling effect of age group – with the youngest workers at

[5] NACE2RevisionCodes: 13,14,28,29,31

[6] NACE2RevisionCode: 25

[7] NACE2RevisionCodes: 41,43

[8] NACE2RevisionCode: 55

[9] NACE2RevisionCodes: 05 – 09

[10] NACE2Revision Code: 68

greatest risk of unemployment first followed by those just starting out in employment and household establishment.

Predictors of being unemployed in 2010

Once again it is *occupation* at the forefront of determining risk of unemployment, but now the list of 'risky' jobs to hold in 2007 has widened to include WHOLESALING[11], PUBLISHING[12] and FISHING[13], in addition to the other occupations that appeared in 2008 and 2009 (Hotels reappears in this year). It is important to recall that the dataset being used to identify these risk factors was constructed in such a way as to include only the first time an individual became unemployed in each of the three years following the crisis in 2007. Hence, the fact that 4 occupations appear in each of the 3 years means that the shedding of jobs in these industries is affecting new people each year. The appearance of WHOLESALING suggests that the decrease in purchasing power of Irish citizens has begun to impact on employment three years after the crisis began. *Age* (including family cycle as a related demographic) continues to be a key risk factor in this year as does *education* – although education drops out of the top three in this year. One new factor appears in the top three and that is *union membership*. Specifically, those who are not members of a union are more prone to become unemployed in this year. This is a very interesting finding, although somewhat difficult to interpret given the delay between the time of the crisis and the appearance of this characteristic as a factor in employment risk.

In terms of the hypothetical 'story' of unemployment risk after an economic crisis, the data suggests that occupations such as construction, manufacturing, and hotels were the leading indicators of risk of unemployment for as long as the recession holds. Age is also a significant risk factor throughout the period with a kind of rolling effect from the youngest workers through to newly formed households. Another 'rolling' effect appeared across the property-related occupations, with the decline in construction followed by a decline in real estate jobs. Higher levels of education and union membership appear to have an inverse relationship with unemployment (i.e., delay unemployment risk), but future research is required to understand which of these is more important.

While the above discussion is not presented as solid evidence to inform policy or practice in the development or implementation of unemployment programmes, it certainly provides some very interesting results that could inform more rigorous investigations into risk factors leading to unemployment after an economic crisis and the likely timing of redundancies. Furthermore, the results and the process leading to these provides some indication of the opportunities and challenges in using existing administrative data for generating useful information without recourse to expensive research programmes or, even more costly new data collection efforts. The value of the findings, once validated, are their potential contribution to the development of more targeted programmes aimed at avoiding unemployment for those at greatest risk,

[11] NACE2RevisionCodes: 45,46

[12] NACE2RevisionCode: 58

[13] NACE2RevisionCodes: 03

thereby lowering both the cost of unemployment and the cost of providing 'broad-brush' supports to those who are at lower risk. In addition, the availability of the data and the analytic tools to perform sophisticated analysis with relative ease suggests that - with some investment in the infrastructure and skills to undertake projects of this type - the public sector could save time and money that would have otherwise been spent on hiring researchers and consultants to perform one-off studies. The opportunity of 'doing more with less' is evident here.

However, the challenges are also apparent. The need to be sworn in as an 'Officer of Statistics' before being allowed to access data seems a significant barrier to expanding the use of administrative data across the public sector and beyond. The anonymisation process seems straightforward, but as it currently stands would put a heavy burden on the CSO if more analyses of this type were contemplated. The lack of common identifiers for individuals across different agencies is also a major problem and one that is widely recognized in Ireland and throughout Europe. Finally, the lack of statistical / analytic skills of the type required to conceive of analyses and carry them out in the public sector is probably the greatest single barrier to engaging in more sophisticated and effective use of administrative data. None of these challenges is insurmountable, but equally none will be met without concerted effort.

References:

Boyle, R. and MacCarthaigh, M. (2011) "Fit for Purpose? Challenges for Irish Public Administration and Priorities for Public Service Reform", *State of the Public Service Series, Research Paper No4*, Dublin: Institute of Public Administration

Department of the Taoiseach (2008), *Transforming Public Services*, Report of the Task Force on the Public Service, Dublin: Stationery Office

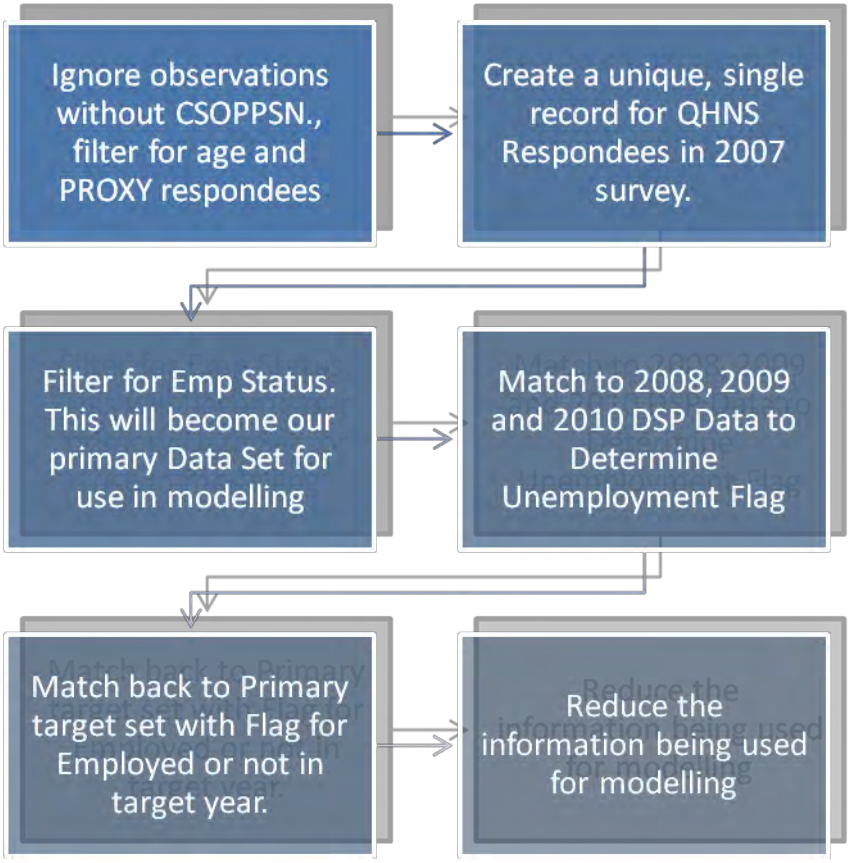
Ljungqvist, L. and Sargent, T.J. (2008),"Two Questions about European Unemployment", *Econometrica*, Vol. 76(1), pp. 1-29

OECD (2008), *Ireland: Towards An Integrated Public Service,* OECD Public Management Reviews, Paris: OECD Publishing

Payne, C. and Payne, J. (2000), "Early Identification of the Long-term Unemployed", *PSI Research Discussion Paper No. 4*, accessed 25 July 2011: http://ideas.repec.org/p/psi/resdis/4.html

Report of the Special Group on Public Service Numbers and Expenditure Programmes (2009), Volume 1, Dublin: Stationery Office

Appendix 1: Process Flow of data preparation post CSO extraction with explanations of each step



Step 1

The first step involved creating a data set that from the 4 quarterly QHNS survey responses in 2007. We filtered these observations to create a superset of observations that had a CSPPPSN, by the CSO's internal PPSN representation. The reason for this is to allow us to match across to the Live register data and P35 data using the CSOPPSN as a unique identifier look up. We then filtered observations for respondents based off their age (using the AgeGroup field) and the Proxy Field (whereby responses to the survey that were for persons under 14 were not permitted).

Step 2

The next step required us to create a unique and single record for respondents irrespective of the number of survey quarters they had appeared in over 2007. E.g. a respondent that was involved in the 2007 survey over Quarters 1,2 and 3 will have 3 records. This causes issues in modelling. Therefore we wanted to create a combined 2007 data set which represented only 1 record per respondent. To do this we matched across all the quarters in 2007 on CSOPPSN. Where duplicates occurred, i.e.,

Before

CSOPPSN	Year	Quarter	Attrib 1
123	2007	Q1	ABC
123	2007	Q2	ABC
456	2007	Q3	MLA
456	2007	Q4	MLA
789	2007	Q1	KMN

We took the most recent quarter's observation for that respondent to create:

After

CSOPPSN	Year	Quarter	Attrib 1
123	2007	Q2	ABC
456	2007	Q4	MLA
789	2007	Q1	KMN

While there is the implication that some data can be lost, taking the latest quarterly response data allowed to take the latest information available for that respondent. This created a record-set of 99,061 observations.

Step 3

At this point we had a unique, single record per respondent data set of QHNS data over 2007. The next step was to create the data set that would contain only respondents who had a principal economic status(PES) of 1 (or In Employment). This was done by simply filtering on this attribute. The reason for using this attribute to define employment is that it allowed us filter on a more binary basis irrespective of whether the respondent was working part-time, contract. In addition the PES also codes for people who are retired, in college. Therefore we felt this was the most appropriate mechanism to find people who are in employment as their principal economic status. (The data set contained 44,073 observations)

Step 4

The next step was to match to the respondents from the data set created in Step 3 to each of the Live Register data sets of 2008, 2009, 2010 using the CSOPPSN as the matching criteria. Where we received a match that record (CSOPPSN) was flagged as being unemployed in that year. In addition we filtered out matches to a respondent that had been unemployed for more that 1 year. i.e. we only created a match to that specific year. The reason for this is for our modeling we only wanted the respondents / records who were made unemployed in that year i.e. not a cumulative set of respondents in 2009 and 2010.

Step 5

This is the final step where we appended our new flag of Unemployed (1) or Employed (0) to the data set created in Step 3. In fact we created 3 variables for clarity, one for each of the years the respondents were first made unemployed.

Step 6

This step involved reducing the amount of variables we are introducing to the modeling process. We deleted from our data sets the attributes of:

- NACERev1Detail codes and Sector Codes
- Occupation Codes
- NaceRev2Sector Codes.

The reason for this was we are using NaceRev2Revision Detail Codes, firstly as the proxy variable representing the above and secondly to understand the effects of more detailed job codes as specified in the NaceRev2Revision Detail Codes.

Appendix 2: Summary of the parameters selected in the creation of the Decision Tree and Regression Models in SAS Enterprise Miner 6.1

Property	Value		Property	Value
Component	**Decision Tree**		Max Depth	8
Assess Measure	Misclassification		Nominal Criterion	Pearson Chi Square
CrossValidation	No		Significance Level	0.2
Bonferroni Adjustment	Y	Whether Bonferroni Adjustment should be applied to p-value assessment of decision splits		
Time for Kass Adjustment	Before	When Bonferroni Adjustment applied to p values		
Leaf Size	5	(smallest no of obs per leaf)		
MaxBranch	2			

Property	Value		Property	Value
Component	**Regression**		Entry Significance Level	0.05
Error	LOGISTIC		Stay Significance Level	0.05
Link Function	LOGIT		Model Selection	Stepwise
Max Steps	None	Specifies Ceiling of steps to be included in model		
Main Effect	Yes	Include all effects available to the model		
SelectionCriteria	Misclassification			
TwoFactorInteractions	No	Whether to include two by two factor interactions		

Appendix 3: Summary of the key parameters and options chosen for evaluating models in SAS Enterprise Miner 6.1 and the results of these tests for the models selected for each year (2008, 2009, 2010)

Options available:

- **Akaike's Information Criterion** — chooses the model with the smallest Akaike's Information Criterion value.
- **Average Squared Error — chooses the model with the smallest average squared error value**.
- **Mean Squared Error** — chooses the model with the smallest mean squared error value.
- **ROC — chooses the model with the greatest area under the ROC curve**.
- **Captured Response** — chooses the model with the greatest captured response values using the decile range that is specified in the Selection Depth property.
- **Gain** — chooses the model with the greatest gain using the decile range that is specified in the Selection Depth property.
- **Gini Coefficient** — chooses the model with the highest Gini coefficient value.
- **Kolmogorov - Smirnov Statistic** — chooses the model with the highest Kolmogorov - Smirnov statistic value.
- **Lift** — chooses the model with the greatest lift using the decile range that is specified in the Selection Depth property.
- **Misclassification Rate — chooses the model with the lowest misclassification rate.**
- **Average Profit/Loss** — chooses the model with the greatest average profit/loss.
- **Percent Response** — chooses the model with the greatest % response.
- **Cumulative Captured Response** — chooses the model with the greatest cumulative % captured response.
- **Cumulative Lift** — chooses the model with the greatest cumulative lift.
- **Cumulative Percent Response** — chooses the model with the greatest cumulative % response.