Problem Outline:



An In-depth analysis oN commute, travel patterns, population explosion and traffic forecast

Version 1.0

April 13, 2018

If you live in a large city or any area where there are lots of commuters on the road, then you are probably familiar with the effects of traffic congestion. What you may not realize is the extent of the effects heavy traffic congestion can have. This gridlock can have a tremendous impact on your personal life, career, your future and even your safety. Finding a solution to traffic congestion could mean a vast improvement in the quality of life in your area.

The first thing many people think of when it comes to congested roadways is the delay. During the morning commute there is additional stress because delays caused by traffic can make people late for work. And at the end of the day, the afternoon rush hour is again a frustrating time because the workday is done and people want to get home to relax, and traffic is preventing it. These delays are the effects most people feel because they are universal to everyone who has to maneuver through congested roads.

A secondary effect of traffic congestion related to delays is the inability to estimate travel times. Those who regularly travel congested areas know approximately how long it usually takes to get through a particular area depending on the time of day or the day of the week. These experienced city drivers have to build in time "just in case" the traffic is bad. This takes away from leisure time and time to do other tasks throughout the day. Also, on a day when the traffic is unusually light the built in extra time may be of no use and the person arrives too early.

The stopping and starting in traffic jams also burns fuel at a higher rate than the smooth rate of travel on the open highway. This increase in fuel consumption costs commuters additional money for fuel and it also contributes to the amount of emissions released by the vehicles. These emissions create air pollution and are related to global warming.

Road rage is a senseless reaction to traffic that is common in congested traffic areas. If someone is not driving as fast as the person behind him thinks he should, or someone cuts in front of someone else it can lead to an incident that is dangerous to the offender and those around him on the road. Road rage often manifests itself as shouting matches on the road, intentional tailgating, retaliatory traffic maneuvers and mostly a lack of attention being paid to the traffic around the people involved. It is basically a temper tantrum by frustrated drivers in traffic.

When you dial emergency and request a police officer, an ambulance or a fire truck and the emergency vehicle is unable to respond in an appropriate amount of time because of traffic congestion which can be a danger to you and your property. Although systems are available that help alleviate the problem by allowing the emergency crews to automatically change the traffic lights to keep the line moving, its high time to alert ourselves regarding the growing nature of traffic.

Aim and Project objectives:

The primary aim of this project is to build a model that forecasts traffic in Galway by taking into account various factors such as population growth, commute patterns and live traffic data from sensors. Apart from these an exploratory analysis is also done on indicators such as population distribution, employment distribution, means of travel, time taken to journey, departure time and car ownership across different counties in Ireland from 2011 to 2016. The project also has a scope for predicting the population explosion in Galway. Also, spatial analysis is used to get a geographical sense of the data, highlighting the distributions and trends while also allowing analysis to be interpreted visually.

Data Understanding:

Once the objectives are clear, next step is to do a thorough analysis on data to gain insights and for the purpose we have followed steps of CRISP – DM methodology in data understanding. This framework helps us in getting familiarized with the data, to know what each variable conveys and examining the quality of the data sets.

This data understanding phase is divided into following subsections:

* Data Collection
* Data Description
* Data Exploration
* Data Quality Verification

## Data Collection:

For the purpose of determining the factors contributing to traffic growth, census data provided by central statistics office popularly known to be SAPS – Small Area Population Statistics has been used. It provides information about economic, demographics, culture, households etc. pertaining to different geographic sub-regions of Ireland – including by county and by electoral division. The SAPS data for years 2006, 2011 and 2016 has been considered for analysis. For 2011 and 2016, all the factors or themes were segregated as tables inside a single CSV file whereas for 2006, each factor forms a CSV and each of them has to be separately downloaded and then all the needed factors has to be combined into one single file. Apart from the SAPS dataset, we also have live traffic data which is taken from National Roads Authority (NRA). Every day traffic detailed to each hour, vehicle type, direction has been recorded by sensors functioning at different routes and they have been uploaded as a CSV file for each route so that users can make use of it. It is also possible to collect the data at different levels of time frame say daily, weekly, monthly, yearly or given time period. One more thing to note is the fact that the data has been collected from the year 2013 to till date. Traffic data prior to 2013 exist but has been inconsistent, in that data for few routes were missing for certain time period. Initially we planned to do traffic prediction for all the routes present in Ireland but since there are lot many routes, the process of collecting data for each of the route for five year time period became tedious and also accounted for many different files. So we decided to do the prediction model for Galway which can then be extended to different counties in Ireland. For Galway, there are five major routes that exist and they are route M6, N17, N18, N59 and N84. For each of these routes, traffic data for last five years (i.e. from 2013) has been collected separately. The traffic data needs a lot of transformation as data is not in the needed structure. So far the traffic data from needed routes have been collected and studied alone. Working with this data set proves to be challenging as it needed more cleaning and transformations.

## Description of DataSet:

Here in our case we got two different data sets – one providing information on the demographics & other factors across counties, other providing data on traffic count. The SAPS data set contains lot many themes from which only certain that are relevant for our analysis has to be taken. Below is the figure showing the relevant fields taken from different themes from SAPS worksheet.



The above fields will be used in analyzing how they influence traffic growth in different counties in Ireland. In addition to the above data set we have traffic data recorded from TII (Traffic Infrastructure Ireland) produced by national roads authority.



The data from CSV is not in tabular form as mentioned above. Instead in a form that needs to be transformed to the above format.

## Data Exploration

Data Exploration was a crucial step before beginning the data analysis which helped us in closely examining the data for each variable. By exploring the dataset, we became familiarized with the dataset, analyzed the data quality issues (structure of the data) and developed initial understanding about what results can be expected from this dataset. Data exploration phase helped us in identifying the research questions and determining the scope of our project more concretely.

In SAPS dataset, there were many themes and many fields. After careful observation of each theme and fields that are contained inside it, we selected certain fields that may be useful in analyzing and building our model. For example themes like Migration & Ethnicity, Language, Communal Establishments, Social Class and Social Economic group, Disability, Internet access doesn’t do anything with influencing traffic and hence they were out of contention. Similarly in traffic count data, we narrowed our focus onto Galway as considering other routes seem tedious work.

## Data Cleansing and Transformation:

1. Transforming the Census Data:

We have two different information needed for the analysis. First one is the census data (SAPS). As we planned to use census data for the years 2011 and 2016, we gathered them separately as they were not in a single CSV file. 2011 and 2016 have all the themes but we were needed to extract only certain themes like population, economic (employment), means of travel, journey times, departure times, car ownership. We used MS Excel to filter out the themes we needed. Also, the county names and ID’s were not consistent among the two census datasets. In 2011, there were 34 listings in the county names whereas for 2016 it was reduced to 31. The problem is because they didn’t followed standard set of county names in the listings. For ex: Cork was given as city and county in 2011 but made into a single term as Cork in 2016. So in order to find this we used Excel to sort and identify the counties that are not matching from 2011 to 2016. Then we fixed 2016 as reference which had aggregated set of counties and based upon that we merged the other fields in census 2011. Also another issue we faced was for 2011 as well as 2016 datasets, car ownership information was not given directly. Instead they had given in the form of number of households with 1 car, 2 cars, 3 cars and 4 cars. We made another field separately to calculate the total number of cars owned. We made use of Excel formulas to do that. In the end we had two files for each of the census years (2011 and 2016) which will be imported into SPSS for further exploratory analysis. Since we decided to do predictive modeling on AADT (Annual Average Daily Traffic) for Galway, the analysis demanded a single consolidated census dataset. Using Excel was tedious as many transformations were needed to be made and the approach was also not scalable. So we made an automation in R Script that would take as many Census data sets and would extract the needed information for Galway, do the transformations like renaming the fields etc. Apart from this we also needed to extrapolate the data for the years 2012, 2013, 2014, 2015 as the census data were unavailable for these years. After studying the growth of factors from census data listed from 1991 till 2016, we found the growth factor to be linear and derived formulas for calculating each of the field being used. Finally all the data were written to a single CSV file and get outputted from the R Script.

1. Transforming the Traffic Data:

On the other hand, the live traffic data was pulled from NRA (National Roads Authority) for 5 different routes (each a separate CSV file) in and around Galway. The routes considered were M06, N17, N18, N59 and N84. In NRA, the traffic data was monitored from the year 2013 to till date and the data prior to 2013 were all maintained through manual entry. So there were deviations in routes being used as well. For Ex, there was no record for traffic in N84 prior 2013 (maybe the route was newly constructed or was not operational before 2013) and records for certain months were missing for other routes. The data set was also unstructured and needed lot of cleansing. We collected all the CSV files (around 30 files) and cleaning it with Excel seemed a tedious work that would consume lot of hours. Also the fact that traffic data files were of different structure; the ones after 2013 following a pattern and prior 2013 following another. So we decided again to automate the process by writing an R Script that would have all the raw data ingested initially, do the transformations by picking the right data we needed (Year, Mean Peak Am, Mean Peak Am Volume, Mean Peak Pm, Mean Peak Pm Volume, Mean AADT). The mean values were calculated based upon the formulas. We also needed to aggregate the traffic volume from each route to get the total volume for Galway city. So the consolidated file after all the merging’s and transformations will be fetched as output of the R Script.

1. Consolidated files:

So here’s the raw census dataset after initial filtering based on theme for each of the years (2011 and 2016)





The consolidated and cleansed Census data from the R Script is shown below:



The live traffic data set that was collected for 6 different routes flowing towards Galway is shown below:

Traffic dataset after 2013 for the considered routes:





Traffic data set prior 2013:



The consolidated and cleansed traffic data from the R Script is shown below:



## Tools Used:

The aim of our project was to conduct meaningful analysis by demonstrating all acquired skills gained during our whole programme. This section provides a summary of important tools and techniques used in data analysis, modelling and visualizations:

|  |  |  |
| --- | --- | --- |
| **Tool** | **Description** | **Project Usage** |
| Microsoft Excel | Spreadsheet Software | For performing initial data loading and merging of different files, Excel has been used. Also for examining the fields and its range of values, it has been used as preliminary tool. |
| SPSS | Statistical family of products for planning, data collection, analysis & reporting | For doing preliminary statistical analysis, to find correlations between factors, normality of the data and various other exploratory tests. |
| R Studio: | Programming language and environment used for statistical computing | Tool that was used for data cleansing, transformation, merging and for building the models using various factors. |
| Tableau | Feature rich visualization application | Used for making interactive visualizations and dashboards. |
| GitHub | Web based hosting service for version control | For maintaining the data, code in cloud and to facilitate team management and source code management. |

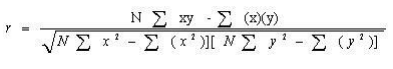
## TECHNIQUES Used:

**Correlations:**

Correlation is a bivariate analysis that measures the strengths of association between two variables and the direction of the relationship. The correlation is one of the most common and useful statistics. Correlations seek to quantify whether there is a relationship between two or more variables. Correlations measures the direction and strength of a relationship between two or more variables. Finding the correlations is very important to identify if there are any underlying relationships between the variables considered for predicting the AADT value.

**Pearson’s Correlation:**

Pearson r correlation is the most widely and frequently used correlation statistic test used to measure the relationship between linearly related variables. The following formula is used to calculate the Pearson r correlation.



R = Pearson r correlation coefficient

N = Number of values in each dataset

∑xy = Sum of the products of paired scores

∑x = Sum of the x scores

∑y = Sum of the y scores

∑x2 = Sum of the squared x scores

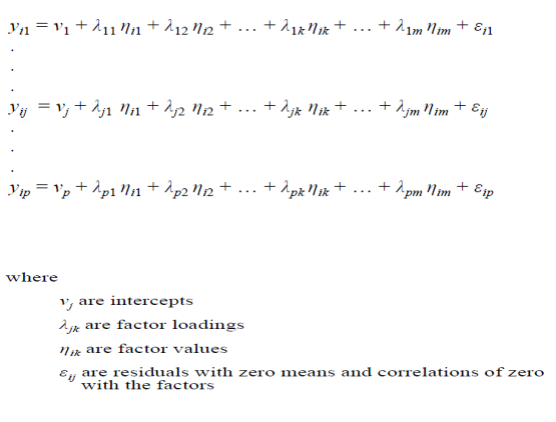
∑y2 = Sum of the squared y scores

**Assumptions:**

For the Pearson r correlation, both variables should be normally distributed. Other assumptions include linearity and homoscedasticity. Linearity assumes a straight-line relationship between the variables involved in the analysis and homoscedasticity assumes that the data is distributed normally about the regression line.

**Factor Analysis:**

Factor analysis is the statistical method to study the dimensionality of the variables. These variables are also referred to as factors. Factors analysis is similar to cluster analysis that groups similar cases. This process is also known as identifying latent variables. Since factor analysis is an exploratory analysis, it doesn’t distinguish between the dependent and independent variables.



Here, factor analysis is mainly used to find the factors that contribute mainly to AADT.

**Multiple Linear Regression (MLR):**

The general purpose of using multiple linear regression is to identify the relationship between several independent and a dependent variable. This process is done by fitting a linear equation to observed data. The population regression line for p explanatory variables x1, x2, ... ,xp is defined to be y = 0 + 1x1 + 2x2 + ... + pxp. This line describes how the mean response y changes with the explanatory variables. The observed values for y vary about their means y and are assumed to have the same standard deviation. The fitted values b0, b1,...,,bp estimate the parameters 0, 1, ..., p of the population regression line.

## Data Analysis:

Identification of the factors:

The important factors that contribute to increase in AADT (traffic volume) needs to be identified and performing Correlation analysis is one of the way. The factors that were used for analysis are:

1. Population Distribution
2. People Employed
3. Unemployed Count
4. People Commuting to Work or School through various modes of transport like Bicycle, Public transport, Motorcycle, Car usage as a driver, Car usage as a passenger, Van, Other modes like lorry etc.
5. Time leaving to Work or School
6. Journey time taken to reach destination
7. Car ownership

Based on the above factors, a research was made to assess how each factor contributed to the traffic growth (AADT).

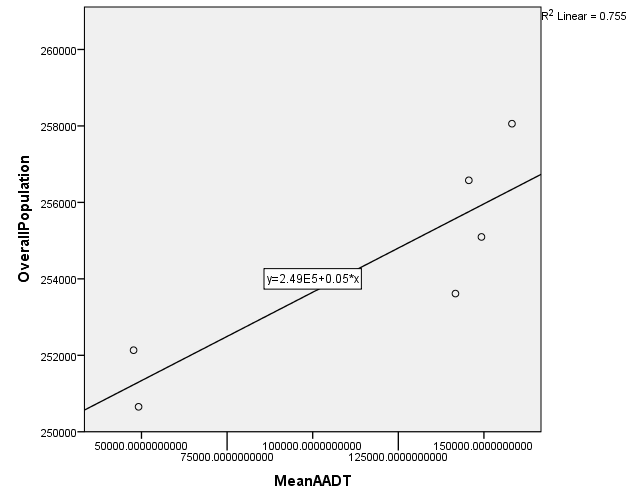
Correlation between the factors/variable:

To compare the correlations between two or more variables, a Pearson’s product moment correlation test can be performed. It was performed to identify if there exists any correlations between the variables considered for the analysis. The results are tabulated as follows:



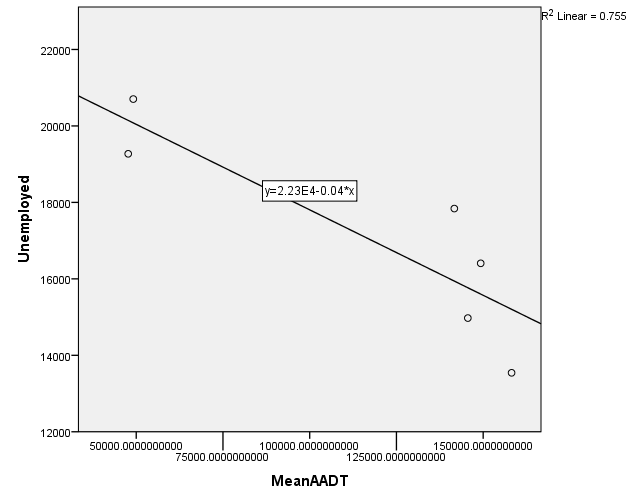
The table shows that factors like population distribution, employment distribution, modes of transport and car ownership are strongly correlated with AADT (.869) and an increase in any of the above factors results in an increase in traffic volume. Also it is important to note that unemployment is negatively correlated with AADT (-.869) meaning increase in joblessness tends to decrease the traffic.

The correlation between the variables can be better visualized using a scatter plot. Here the plot shows how population increase contributing to traffic increase.



It is also observed that the trend remains the same for other factors like employment distribution, modes of transport, car ownership. The reason is the linearity of the data between the factors.

Similarly, there exists a negative correlation between unemployment and AADT which can be visualized as below:



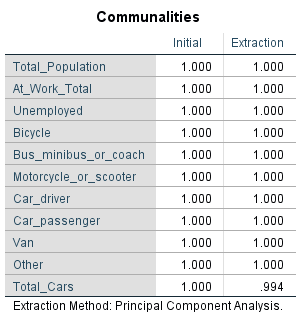
Factor Analysis:

The key concept of factor analysis is that the multiple observed variables have similar patterns of responses because they are all related to a latent. Eigen value is a measure of how much the variance of the observed variables a factor explains. Any factor with an Eigen value >1 means a greater variance than a single observed variable.

The variables that were considered for the factor analysis are:

1. Population Distribution
2. Employment Distribution
3. Unemployed Population
4. Modes of transport
5. Car ownership

The Eigen values for the factors and the variance levels were calculated initially. The results are shown below.





In order to understand the relationship between the Eigen values and the factors, a scree plot is represented as below:



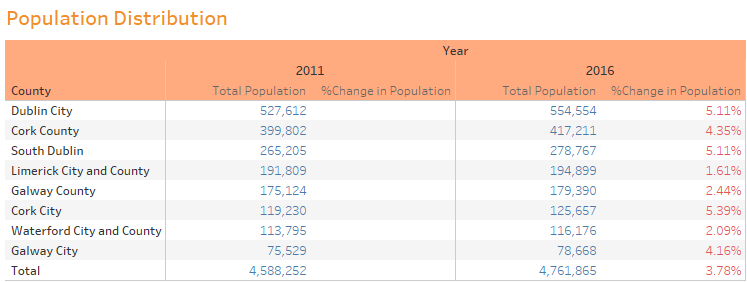
The above plot shows that total population to be a predominant factor and states that the other factors are ignorable. This can be related to the shortcoming of having less data as the system couldn’t find proper relationship with other factors and hence the curve steeps down. In order to demonstrate that the factors considered influence the traffic volume, following visualizations and findings were made.

## Further Analysis:

A detailed study on the following factors were made to demonstrate their effect on traffic volume growth that are key to traffic planning.

1. Population Distribution:

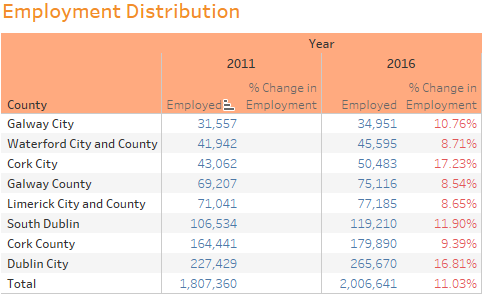
The distribution of population is a critical determinant of travel behavior. In particular, the location of population relative to key services such as work, education, retail and leisure determines the demand for travel and distances travelled. In turn this has a critical impact on people’s choice of mode – i.e. the more remote people live from public transport and destinations served by public transport the more likely they are to drive.



The table shows that the population increase across the nation has been 3.78% from 2011 to 2015 and interestingly Galway City accounted for 4.16% increase. Cork city recorded the largest growth with 5.39% while Dublin showing 5.11% positive growth.

1. Employment Distribution:

The distribution of employment is a key determinant of travel patterns – particularly in the morning peak period when most people travel to work. It would be interesting to note how the employment distribution has been since Census 2006 was undertaken as Ireland was coming to the end of an economic boom, while the 2011 census is the first census in over 15 years to have been undertaken in the midst of an economic downturn and 2016 again being a fruitful year in terms of employment opportunities. The changes in employment levels and the distribution of these changes between the census years is a key indicator of the impacts of the economic growth.

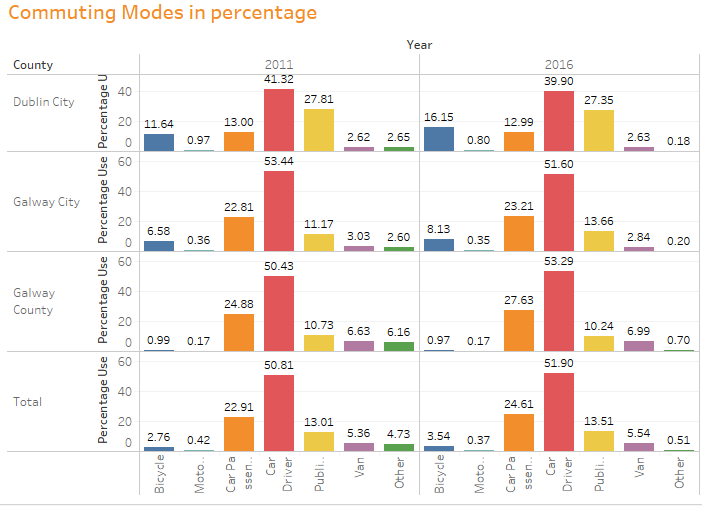


It can be clearly seen that the employment rate has grown thick from 2011 to 2016 with a national increase of 11.03% and Galway city showing considerable increase of 10.76%. Cork and Dublin records the major increase with 17.23% and 16.81%.

1. Means of travel:

While the distribution of population and employment are critical determinants of overall travel demand, the means of travel people choose is an important measure of travel behavior. Means of travel is affected by a number of factors, including the location of development, general economic conditions, availability of public transport, changes in fuel costs and public transport fares and other transport policy interventions. The economic boom between 1994 and 2006 had all the classic impacts on travel behavior in Ireland as have been experienced in other developed economies in the world – i.e. increases in car ownership and in car use, increased traffic congestion in cities and a general decline in the use of public transport and other sustainable modes (e.g. walking and cycling). However, while the national trend was one of increased car use and car dependency, this trend was significantly reversed in the core of the major urban centers by the introduction of parking controls and by investment and improvements in public transport and other sustainable modes.

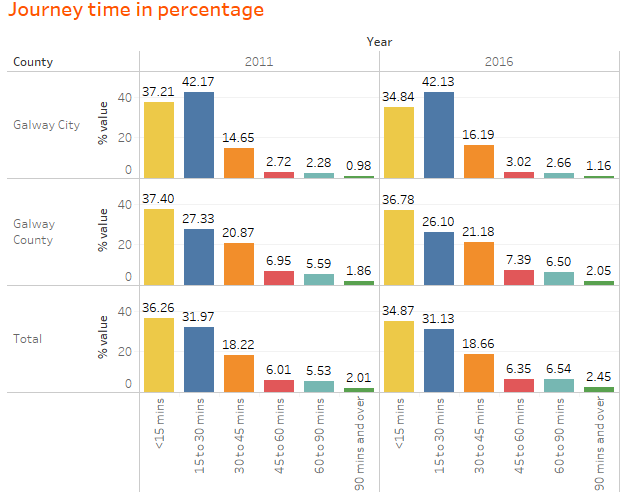
An analysis of the means of travel to work and education in the 2016 Census and comparison with equivalent data from earlier census years gives a good indication of the impacts of the recent economic changes and other factors on people’s travel behavior.



Both tables show similar trends at the Galway city and national level with an increase in mode share for car and reductions in walking and travel by bus. However, the only mode to increase in absolute terms was cycling. There was an increase of around 2% in cycling trips to work nationally as well as in Galway city with Dublin recording maximum of 5% increase in cycling. This increase in cycling has been influenced by a number of factors in recent years, including the increases in fuel costs and in public transport fares and investment in cycle networks in urban centers. The success of the bikes scheme has also had a major impact on attitudes to cycling – particularly in Galway and Dublin city. Of note are the trends in mode share for car drivers and car passengers. While the mode share for car as passenger has increased both in the Galway and nationally, the mode share for car as driver has declined. This reflects the increase in use of carpooling for trips to work. A slight increase in use of public transport can also be seen from the graph which is a thing to welcome.

1. Journey time:

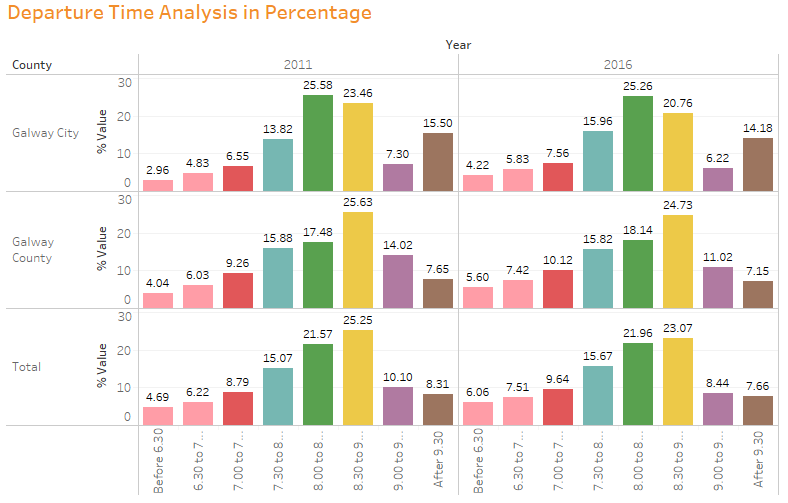
The economic boom between the mid 1990’s and 2006 had the characteristic impact of increasing overall travel demand and traffic congestion levels – in particular in urban areas. This in turn had an impact on journey times to work in particular. An analysis of journey to work times from Census 2011 and 2016 is useful to reveal if the trend up to 2006 has been halted or reversed in response to the economic changes.



It can be clearly noted that the people taking journey times 30 to 45 mins, 45 to 60 mins, 60 to 90 mins and 90 mins over has grown in number both in Galway city as well as nationally and can be associated with proportional increase in traffic congestion.

1. Departure time:

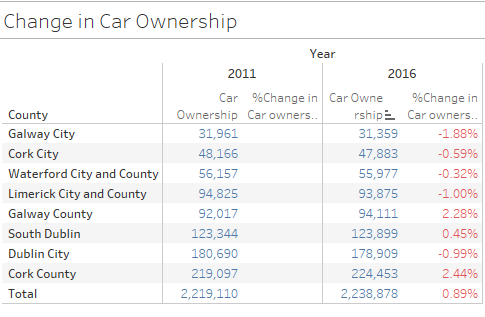
Increased traffic congestion in the economic boom years made for longer journey times and caused commuters in urban areas in particular to depart earlier to be certain of arriving on time for work. The phenomenon of peak spreading was particularly evident nationally as revealed by comparison of travel to work departure times in the 2002 and 2006 census years. Analysis of departure times for travel to work from Census 2011 & 2016 is useful to reveal if the trends in peak spreading have altered or reversed in the past five years.



The plot shows that people have started to leave earlier to avoid stuck in traffic. It is also interesting to note that the national peak hour for travel to work or school has been 8.30 to 9.00 whereas Galway city differs in it and has 8.00 to 8.30 as peak time.

1. Car ownership:

Levels of car ownership have a critical impact on people’s travel behavior – in particular on the number of trips they will make and the means of travel they will choose. Prior to the economic boom in the early 1990’s, Ireland had one of the lowest levels of car ownership in the EU with one in four persons in the state owning a car. As has been the case in most developed economies worldwide, the economic boom in Ireland lead to huge increases in car ownership to levels that are now on a par with many of our EU partners. Below is the plot that shows how the car ownership has changed from 2011 and 2016.

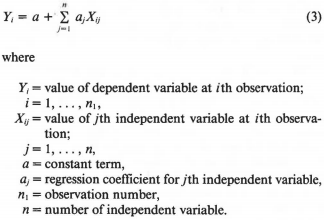


Though most counties didn’t show a major change or reduced slightly, Galway County showed a 2.28% increase in car ownership which reflected in increase in traffic volume over the period.

## Prediction Model:

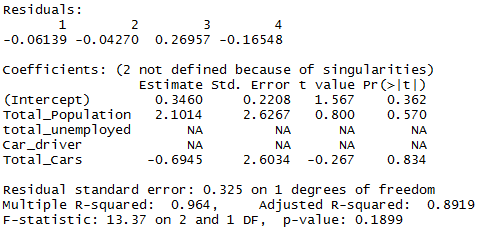
Estimates of future traffic can be obtained by two very different methods: trend projections and forecasts. Analysts can modify extrapolated trends based on their experience and knowledge of the route, state, or region. With trend projections only traffic data are being dealt with; in forecasting techniques, however, a relationship between traffic and explanatory factors must be established. Traffic forecasting techniques therefore are also concerned with knowing the future values of economic and other measures or indicators of person and travel.

Multiple Linear regression was used to identify factors that best estimated AADT and their respective elasticities. The factors used were population, number of households, automobile ownership, and employment. Here the response variable (dependent variable) used is AADT (Average Annual Daily Traffic) which was calculated based on average number of vehicles that crosses a particular route for a period of one year. Since traffic estimation is being done for Galway, AADT measures of 5 different routes linking to Galway was considered. Ireland has been recording the daily traffic flow using sensors from the year 2013 for the routes where the system is in place. A manual entry of traffic flows across a route was in place before 2013. The manual recording however was very inconsistent, in that records for many routes weren’t added. This made us to consider the dataset for a period of 6 years from 2011 to 2016. Independent variables like population, unemployment, car ownership, modes of commute taken from census data needs to be extrapolated for the years in between 2011 and 2016. After checking the linear growth rate examining the census records from 1991 till 2006, the values of these variables are extrapolated. With a record set of just 6, the model has been made using linear regression. Below is the mathematical equation of the linear form:



Training set consisted of first 4 records and test set contained the remaining 2 records. In order to uniform the values of different variables, normalization has been made and training set was given to the linear model.

Below is the summary results of MLR model built using the factors Population, Unemployment, Commute Using Car as Driver, Car ownership:



The predicted AADT values for 2015 and 2016 is 87000 and 94000 respectively whereas the actual AADT is 77000 and 80000. Although the variance is huge, it has reasons for it. Below are the challenges we faced in collecting the data and building the model.

Challenges:

Although the approach in forecasting makes sense, the model has more deviations because there were challenges in the data set. The response variable AADT (Average Annual Daily Traffic) depends on the traffic data being collected. Since the data is being collected only from the year 2013 for all the routes and the traffic count for the years before 2010 were not maintained properly, this made us to consider the AADT for the years 2011 to till date. In order to get the factors that the AADT would actually depend on, we considered demographic and economic criteria like Population, households, employment, commute modes, car ownership. These factors were fetched from the census data. Although the census data existed from the year 1991 till 2016, as the AADT values were known only from 2011, we decided to consider census 2011 and 2016. Also the census data needs to be extrapolated for the years 2012, 2013, 2014 and 2015. Here we considered the growth factor of each factor to be linear after examining the trends from 1991 till 2006. With a record set of 6, it’s difficult to get the relation between factors and also due to multi collinearity, coefficients of certain important factors became NA and has to be excluded from the model. This also contributed to such a variance in the actual and predicted values. It is assured that the model would have performed better if the record set is considerably larger.