Road Traffic Correlations with Economic Variables: The Big Data Perspective

Kilian Ros
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
k.m.ros@student.utwente.nl

ABSTRACT
Road traffic demand is an interesting topic of research for a wide variety of policymakers. Traffic elasticity, which shows the responsiveness of traffic intensity to other variables, can predict future demand. Existing research on this topic mainly uses survey and panel data. This paper uses big data from the National Database of Road Traffic in the Netherlands to study the correlation between road traffic intensity in different vehicle categories and several economic variables. The purpose of this paper is to extend earlier research results by using a new, different and very comprehensive data source compared to earlier studies. This study shows that there exist correlations between road traffic intensity and multiple economic variables. Most notable are the correlations between total traffic intensity and income (0.74), small vehicle traffic intensity and income (0.74) and the correlation between large vehicle traffic intensity and GDP (0.70). It also concludes that the road traffic in the Netherlands is increasing during the research period of four recent years.

Keywords
Economic Variables, Fuel Price, Income, Road Traffic, Intensity, Volume, Elasticity, Demand, Peak Car, Big Data

1. INTRODUCTION
Road traffic demand has been an interesting topic of research from an economic, as well as a political and environmental point of view. Quantifying this demand can have great influence on the decision making process of various policymakers. Think of fuel pricing, political, transport planning, and environmental policies.

Existing research on road traffic elasticity and correlation with variables mainly uses survey and panel data. While these data sources can provide a rather good estimation of the actual situation in the real world, they are not based on actual measured traffic data. This paper will use open data of the National Database of Road Traffic (NDW) in the Netherlands [3]. It provides actual traffic data collected by more than 27,000 sensors located across the Dutch highway network. The sensors are capable of measuring traffic intensity, which is the amount of vehicles passing by a sensor, the average speed of vehicles passing by a sensor, the realized and estimated travel time, which is measured between two sensors, and also the vehicle category, which is based on the length of a certain vehicle. We are only interested in the traffic intensity sensors that are capable of distinguishing different vehicle categories. The selection of sensors will be discussed in the methodology section.

The purpose of this paper is to extend earlier research results by using a new, different and very comprehensive data source compared to earlier studies. By using recent data from the last decade, this paper will provide up-to-date numbers about the correlation between road traffic intensity and several economic variables. The paper will focus on the available data from the Netherlands.

2. RESEARCH QUESTIONS
This paper addresses the following research questions:

RQ1 What are interesting economic variables for testing on correlation regarding road traffic intensity?

RQ2 What is the correlation between road traffic intensity and these economic variables in the Netherlands?

RQ3 How do the results differ when looking at distinct vehicle types?

3. RELATED WORK
The research in elasticity of road traffic with respect to the economic variables fuel price and income has shown that there exists significant correlations. In 2004 two parallel "blind" literature reviews on this topic were conducted by Goodwin [8] and Graham [10]. Both reviews included 69 historical works of about 15 different countries, including The Netherlands, covering a period from 1929 to 1991. Main findings which are meaningful to this paper are:

1. If the real price of fuel rises by 10% and stays at that level: The volume of traffic will fall by roundly 1% within about a year, building up to a reduction of about 3% in the longer run (about 5 years).

2. If the real income goes up by 10%, the following occurs: The volume of traffic will grow by 2% within a year and about 5% in the longer run.

Graham adds to these results that there was no sufficient information to also verify this for freight traffic. This means that the above numbers only apply to smaller vehicles like cars and motorcycles.

Next to these results, there are also researchers who advocate that recent developments in road traffic intensity...
cannot be explained by looking at the traditional economic variables such as fuel prices and income [13]. Puentes and Tomer showed that vehicle use has been declining since 2007. This phenomenon was later summarized to the term 'Peak Car'. "Peak car is the debate about whether the long dominant growth in car use specifically has come to an end, is nearing an end, or is turning down, or is only temporarily interrupted" [9].

Research of Millard-Ball and Schipper concludes that because of the 'Peak Car' phenomenon, traditional factors such as income and fuel price "can only provide a partial information of the car use trend" [12]. This research does not include the Netherlands, but it does include other Western European countries like the UK, Sweden, France and Germany.

In contrast to this, recent research of Bastian proves that the economic variables GDP and fuel price are still sufficient to explain the observed trends in car traffic. This holds for the United Kingdom, Sweden, France and to a slightly less extent Germany [7].

Among the related work there are different methods of measuring road traffic intensity. Goodwin and Graham review many earlier works, in which fuel consumption is the most common measure. Puentes and Tomer use vehicle miles traveled as traffic intensity measure. Millard-Ball and Schipper and Bastian both use the same traffic intensity measure: the amount of vehicle kilometers traveled per person. These measures are different from the ones used in this research. We will use the sum off all traffic sensor counts as traffic intensity measure. This measure should not be interpreted as a count of unique vehicles on the road, but more as an indirect estimation of the amount of vehicles on the road combined with the distance that they cover.

4. ECONOMIC VARIABLES

The related works mainly researched economic variables like fuel price, income and the Gross Domestic Product (GDP) to evaluate traffic elasticity. Therefore, it is only logical for this research to also include these variables to be able to make a valid comparison.

Daily prices of three different fuel types were available, namely Euro95 (gasoline), diesel and LPG (gas). Table 1 shows the distribution of fuel consumption for road traffic in the Netherlands. LPG only has a small part in the total consumption and is therefore not interesting to explore any further. The distribution of Euro95 and diesel remains relatively constant over time, which is necessary in order to come up with valid results. Sudden demand changes would have effect on the fuel prices, resulting in a biased research to correlation and elasticity.

Table 1. Distribution of fuel consumption for road traffic in The Netherlands

<table>
<thead>
<tr>
<th>Category</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016 Q1</th>
<th>2016 Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPG</td>
<td>4%</td>
<td>3%</td>
<td>3%</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>Euro95</td>
<td>40%</td>
<td>42%</td>
<td>42%</td>
<td>42%</td>
<td>43%</td>
</tr>
<tr>
<td>Diesel</td>
<td>56%</td>
<td>55%</td>
<td>55%</td>
<td>56%</td>
<td>54%</td>
</tr>
</tbody>
</table>

For the income variable we take the gross disposable income for households in the country. The gross disposable income is the amount of money that households have available for spending and saving after income taxes have been accounted for. GDP is the size of the economy measured in million euros. Both datasets are available in quarterly intervals. Next to these variables there are also other economic variables that might have a relationship to the intensity of road traffic. First of all this research will also take inflation into account. The inflation rate is a measure of the change of the consumer price index (CPI) relative to the same period one year earlier. The CPI represents the price change of a 'bucket' of goods and services that are common for all Dutch households.

Secondly unemployment will also be considered an economic variable which could influence the road traffic intensity. With a lot of commuter traffic in the Netherlands the assumption is that when more people are unemployed, road traffic intensity will decrease. Purchasing power was also considered an economic variable that could be related to intensity, but due to the yearly frequency of the data, this variable would not contain enough data points for detailed evaluation. Table 2 provides an overview of the selected economic variables, their frequency and the corresponding data source.

Table 2. Economic variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel price</td>
<td>Actual daily pump prices of Diesel fuel and Euro95 gasoline.</td>
<td>Daily</td>
</tr>
<tr>
<td>Inflation</td>
<td>Increase in the general price level of goods and services in an economy over a period of time.</td>
<td>Monthly</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>Ratio of the unemployed to the working age population.</td>
<td>Monthly</td>
</tr>
<tr>
<td>Income</td>
<td>Gross disposable income of households in The Netherlands.</td>
<td>Quarterly</td>
</tr>
<tr>
<td>GDP</td>
<td>Gross domestic product of The Netherlands.</td>
<td>Quarterly</td>
</tr>
</tbody>
</table>

5. METHOD OF RESEARCH

5.1 Data Acquisition

After determining the interesting economic variables and obtaining the data, it is necessary to obtain the corresponding traffic intensity data. For this data it is important that the set of measuring sensors is identical and active during the chosen time period of four years to ensure comparable results.

Table 3. Vehicle type distinction in meter [4]

<table>
<thead>
<tr>
<th>Cat</th>
<th>Description</th>
<th>Length-interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>motorcycle, van, passenger car</td>
<td>&lt;5.60</td>
</tr>
<tr>
<td>2</td>
<td>inarticulate truck, inarticulate bus</td>
<td>≥5.60≤12.20</td>
</tr>
<tr>
<td>3</td>
<td>articulated truck</td>
<td>&gt;12.20</td>
</tr>
</tbody>
</table>

During the years 2013, 2014, 2015 and 2016 there were 4955 traffic intensity sensors available that were also capable of distinguishing between three different vehicle categories. The locations of these sensors are plotted on the left map of Figure 1. The distinction between three different vehicle categories is determined according to the length intervals in Table 3. The total traffic intensity count is not exactly the sum of the three categories. This is caused by...
the accuracy of the sensors. When a sensor cannot allocate a vehicle in a particular category, it is included in the total vehicle count but not in any of the category counts. The traffic intensity data is available per minute, but because of the daily frequency of the economic variable fuel price, which is the highest frequency among variables, the intensity data is also aggregated per day. The complete dataset of all sensors is collected by using the NDW historical database. The data is downloaded as two CSV files per month. Every month contains about 1.25 million rows of data, which add up to a total amount of more than 50 million data rows to analyze. The total file size of the complete dataset is around 25GB. Table 4 provides a description of the most important data fields in the dataset that are used in this research. With these data fields we are able to get the average amount of vehicles passing by a particular sensor per hour on a certain day. The specificVehicleCharacteristics field makes it possible to distinguish between the three different vehicle categories and the dataError field gives indicates whether an error has occurred and no measurement was done.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>measurementSiteReference</td>
<td>Distinct sensor ID</td>
</tr>
<tr>
<td>periodStart</td>
<td>Start of the measurement day</td>
</tr>
<tr>
<td>periodEnd</td>
<td>End of the measurement day</td>
</tr>
<tr>
<td>dataError</td>
<td>Error monitoring. 1 if error occurred, 0 if not</td>
</tr>
<tr>
<td>avgVehicleFlow</td>
<td>Average number of cars passing per hour</td>
</tr>
<tr>
<td>specificVehicleCharacteristics</td>
<td>Description of the measured vehicle category</td>
</tr>
</tbody>
</table>

5.2 Data Filtering

In order to obtain valid results, the data is analyzed on errors. The analysis revealed that the average error percentage per month is around 7% and that there are outliers with the maximum being around 13%. This means that from all potential measurement points in this month there were 13% that did not produce any data output because some kind of error occurred.

When looking at the small percentages of changes in the related work it can not be justified to work with data that contains this many errors and these outliers. Therefore the data has to be filtered so that it will not influence the results.

The first approach to filter out errors is to see how many sensors remain after ignoring all those sensors that have ever produced an error. This method would leave a subset of around a thousand sensors. At first this looks like a fair amount to work with, but when plotting the sensor locations on a map like Figure 1 it turns out that this subset is poorly distributed across the country and therefore not useful.

The second approach filters out the most erroneous sensors of the outlier months with the intention to bring the outliers back towards the overall average of errors. This turns out not to be useful because these particular sensors also produce errors in all other months which are then also filtered out. This causes the error percentage to decrease but the outliers still remain.

The final approach sorts the sensors based on the total amount of errors that they have produced and filters out 2455 sensors with the most errors so that a subset of 2500 sensors remains. This number is established by trial and error. The subset still contains two months that are considered outliers: January 2013 and September 2016. Because these months lay at the beginning and end of the time interval of this research and the high frequency of the
intensity and variables data, we choose to exclude them, and also the remaining months after September 2016, from the research. This means that the time period is adjusted to range from February 2013 till August 2016. For this time period the data contains an average error percentage of 0.11% without outliers and a reasonably good coverage of the country. The locations of the remaining sensors are plotted on the right map of Figure 1.

5.3 Data Processing
For all of the 2500 sensors the data is summed up per vehicle category per day. This means that the number we are actually working with is more like a volume of traffic intensity, which is harder to interpret. Dividing the number by 2500 sensors and multiplying it by 24 hours would give a more recognizable number because it would display the average amount of vehicles per day per sensor. We choose for the simple sum method because it is not about interpreting the number but about comparing the numbers to each other.

The economic variables other than fuel price have a different frequency and therefore the daily intensity data must be adjusted accordingly. For the monthly variables inflation and unemployment, the traffic intensity number is the average of the daily numbers for the particular month. For the quarterly variables GDP and income, the traffic intensity is the average of the monthly numbers for the particular quarter. Because the first quarter of 2013 and the 3rd quarter of 2016 are both missing a month of data, these are ignored.

5.4 Correlation Method
The most common method to calculate correlation coefficients is Pearson’s product-moment correlation. Pearson’s correlation coefficient is a measure of the strength of the linear relationship between two variables that follow a normal distribution [11].

In order to use this correlation coefficient, the data series have to be checked on the assumption of normality. To do so the Shapiro-Wilk test is performed on all the variables present in the data set. The corresponding q-q plots are also analyzed.

If one of the two variables used in a correlation calculation is not normally distributed, then the Pearson’s correlation coefficient should not be used. In this case Spearman’s rank correlation will be used. "Spearman’s rank correlation coefficient is a nonparametric (distribution-free) rank statistic proposed as a measure of the strength of the association between two variables. It is a measure of a monotone association that is used when the distribution of data makes Pearson’s correlation coefficient undesirable or misleading." [11] Table 5 shows an overview of the distribution of a variable and the correlation coefficient method to use.

5.5 Elasticity Method
We define the elasticity of road traffic $e_t$ as a measure to show the responsiveness of the road traffic intensity $T$ to a change in economic variable $V$. More precisely, it gives the percentage change in road traffic intensity in response to a one percent change in the used variable. We prefer arc elasticity over point elasticity because it takes an average which makes it independent from which data point is taken first in the calculation. The formula for the coefficient of road traffic elasticity with respect to a variable is:

$$e_t = \frac{dT}{dV} = \frac{T_2 - T_1}{V_2 - V_1} \frac{(T_2 + T_1)}{(V_2 + V_1)/2}$$

The elasticity coefficients are calculated over the first three years, starting in February. This causes that the quarterly economic variables do not cover exactly the same time period but are off by one month.

6. RESULTS
6.1 Road Traffic Intensity
Figure 2 shows the progression of the total traffic intensity during the period of research. When looking at the graph,
there are a few things that stand out. First of all, we see a clear pattern during a period of a year. From January till June there is a clear upward trend where the first low intensity period in January might be explained by the fact that many working people still enjoy their holidays. In July and August we see a heavy drop in traffic intensity which can logically be explained by the summer holidays in the Netherlands. Most of the people go on holidays during these two months. After that we see an increase in the months September and October which is expected because people go back to work. Then at the end of the year in the months November and December we observe a heavy drop, which is expected in December because of the holidays at the end of this month.

Figure 3 shows the progression of the traffic intensity between the three different vehicle categories. Notice that the left y-axis, which represents Cat1, is of a different scale then the right y-axis. This is done to improve visibility. Roughly 86% of the intensity is generated by Cat1; Cat2 and Cat3 are responsible for 6% and 8% respectively. We practically see the same peaks and drops between the different categories, although one could argue that the standard deviation of Cat3 is higher than the other two categories.

A remarkable fact is that traffic intensity is increasing over the years in all vehicle categories. This is in contrast to the 'Peak Car' phenomenon which states that road traffic has reached its peak and is now decreasing.
### Table 6. Correlations between traffic intensity and economic variables

<table>
<thead>
<tr>
<th></th>
<th>Total Intensity</th>
<th>Cat1 Intensity</th>
<th>Cat2 Intensity</th>
<th>Cat3 Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euro95</td>
<td>-0.12</td>
<td>-0.14</td>
<td>-0.14</td>
<td>-0.14</td>
</tr>
<tr>
<td>Diesel</td>
<td>-0.16</td>
<td>-0.17</td>
<td>-0.09</td>
<td>-0.15</td>
</tr>
<tr>
<td>Inflation</td>
<td>-0.42</td>
<td>-0.43</td>
<td>-0.19</td>
<td>-0.43</td>
</tr>
<tr>
<td>Unemployment</td>
<td>-0.47</td>
<td>-0.41</td>
<td>-0.55</td>
<td>-0.22</td>
</tr>
<tr>
<td>Income</td>
<td>0.74</td>
<td>0.74</td>
<td>0.45</td>
<td>0.65</td>
</tr>
<tr>
<td>GDP</td>
<td>0.54</td>
<td>0.63</td>
<td>-0.01</td>
<td>0.70</td>
</tr>
</tbody>
</table>

White indicates p value < 0.01
Light grey indicates p value < 0.05.
Dark grey indicates p value > 0.05.

An interesting correlation is the one with Cat2 intensity, it is considerably higher than the other categories. Although we can only hypothesize about this difference, we think that it might be caused by a large drop in employment within the construction industry. During the years 2013-2016 approximately 25,000 jobs were lost in this sector, which is a drop of about 6% [5]. We assume that this sector is largely responsible for driving vehicles in this category because they drive extended vans or carry trailers with them to a construction site.

Figure 4 visualizes the above mentioned correlation coefficients. The plotted figures are scatter plots with the traffic intensity category on the x-axis and the economic variable on the y-axis. The blue line is the linear regression line that best fits the data points. This line indicates the direction of the relationship between the traffic intensity and the economic variable. For the variables Income and GDP we see an upward trend which corresponds to the positive correlation coefficients, for the unemployment variable we see descending lines which correspond to the negative correlation coefficients.

Noticeable is the fact that we did not find any correlation between the fuel prices and traffic intensity. We think that this is caused by the daily frequency of our data. People tend to make travel plans in advance. They are not very likely to change their plans when fuel price changes all of a sudden. The same goes for freight traffic, their routes are mostly planned in advance and sudden changes in fuel price will not cause them to be canceled. Even when we look at monthly aggregated fuel price and intensity data, we see correlation coefficients which are in the lower part of a low correlation interpretation.

In general, we don’t see very large differences in correlation coefficients when looking at the different vehicle categories. Some of the variables show a difference between the different categories, with the largest difference being between Cat1 and Cat2 regarding unemployment. For the total intensity and Cat1 intensity it is logical that there are not a lot of differences since Cat1 is responsible for roughly 86% of the counted vehicles within the total traffic intensity. For Cat 2 and Cat 3 it is unfortunate that a lot of correlation coefficients are not statistically significant, this makes it difficult to completely answer RQ3.

### Table 7. Elasticities of road traffic intensity with economic variables

<table>
<thead>
<tr>
<th></th>
<th>Total Intensity</th>
<th>Cat1 Intensity</th>
<th>Cat2 Intensity</th>
<th>Cat3 Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euro95</td>
<td>-0.51</td>
<td>-0.52</td>
<td>-0.44</td>
<td>-0.60</td>
</tr>
<tr>
<td>Diesel</td>
<td>-0.36</td>
<td>-0.37</td>
<td>-0.31</td>
<td>-0.42</td>
</tr>
<tr>
<td>Inflation</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.04</td>
<td>-0.06</td>
</tr>
<tr>
<td>Unemployment</td>
<td>-0.75</td>
<td>-0.77</td>
<td>-0.66</td>
<td>-0.88</td>
</tr>
<tr>
<td>Income</td>
<td>1.63</td>
<td>1.68</td>
<td>1.43</td>
<td>1.93</td>
</tr>
<tr>
<td>GDP</td>
<td>1.52</td>
<td>1.57</td>
<td>1.34</td>
<td>1.80</td>
</tr>
</tbody>
</table>

#### 6.3 Elasticities

Table 7 presents the elasticity coefficients per couple of variables. As mentioned in the methodology, the elasticity is calculated over a period of three years. An interpretation for the -0.51 elasticity coefficient of total traffic intensity with Euro95 fuel would be that for every 10% increase in fuel price, the traffic intensity would decrease by 5.1%.

When looking at the elasticities of the Euro95 and Diesel fuel prices, we can say that they correspond quite a bit with the findings of Goodwin and Graham. Also the income elasticity matches the one mentioned in their work. Both of the elasticity coefficients that we found are slightly higher than the ones described in their work.

Remarkable is that when we look at the daily fuel price data, we do not find any correlation. But when we calculate the elasticity coefficients, we notice a number that indicates a certain relationship. In the correlation results
we already addressed this to the fact that people do not change their plans when prices change all of a sudden. The elasticity is calculated over a much longer period and is therefore not all of a sudden.

We do see different elasticities among the different vehicle categories. For all variables we see that the elasticity coefficients in Cat3 are higher than they are in the other categories, this would indicate that freight traffic is more sensitive to changes in these economic variables than other vehicle categories.

7. DISCUSSION

We are aware of the fact that the set of road traffic intensity sensors which are used in this research are not optimally distributed across the country and that some road traffic movements might take place without a sensor noticing it. Also movements at some particular roads will be counted multiple times. The filtering of sensors, which created less sensor density in the middle of the country, was unavoidable because we think that high error percentages in the measurements would cause a greater bias than the slightly uneven distribution of sensors. The goal of looking at road traffic intensity like this is to come as close as possible to the real traffic intensity number. We think that the used set of sensors was capable enough to notice meaningful changes in the road traffic intensity.

Furthermore we think that the use of elasticity coefficients is not the best measure of a relationship when there are a lot of data points involved in the dataset, like for example our daily fuel prices. This is because it does not clarify anything about the movements of the data in between the starting- and ending point of measuring. For this reason it might be more appropriate on the longer run than it is on a shorter period of time. With this dataset we can only calculate a three-year elasticity, which might be very well affected by other factors than the one economic variable that we are using to calculate the elasticity. We think that it is too simplistic to say that traffic intensity is directly related to an economic variable because traffic intensity can be affected by much more factors and variables than solely one. Therefore we think that constructing a model which includes multiple variables, like Bastian [7] did, is much more appropriate than simply looking at elasticities that take only one particular economic variable into account.

It is also worth mentioning that we are aware of the fact that some of the economic variables that are explored during this research are not always independent from each other. For example GDP and gross disposable income are not likely to change their plans when a sudden price change occurs.

Furthermore, the elasticities found in this research correspond to that of earlier research, but when datasets are large and contain many data points we are not convinced that elasticity is the best measure of a relationship between an economic variable and road traffic intensity because it does not clarify anything about the movements in between the starting- and ending point of measuring.

Lastly, we found only a few small differences in correlations when looking at the different vehicle categories. It is hard to conclude that there are no large differences between the different vehicle categories because of the insignificance of some correlation coefficients. For the elasticity coefficients we did find sizable differences between the different vehicle categories, most remarkable is the fact that vehicle category three, large trucks, has a higher elasticity coefficient with every economic variable than any other vehicle category.

9. ACKNOWLEDGEMENTS

The author wants to thank the NDW for allowing access to the historical database which is required to obtain the necessary road traffic intensity data and also for the assistance in determining a set of intensity sensors that is identical and active during the chosen time period.

10. REFERENCES

[11] J. Hauke and K. Tomasz. Comparison of values of Pearson’s and Spearman’s correlation coefficients on Cat3 traffic intensity and GDP (0.70). Not finding any correlation between traffic intensity and fuel prices was unexpected but can be explained by the fact that the data was of a daily frequency and that people plan ahead and are not likely to change their plans when a sudden price change occurs.
the same sets of data. *Quaestiones Geographicae*, 30(2):87–93, 2011.
