Bus Analytics - Data Description

Wolfgang Garn

The Surrey Business School, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom

Wolfgang Garn[1]

Department of Business Transformation, University of Surrey, United Kingdom

1. Data Description

The aim is to predict passengers [1] for new and existing routes. This will be based on the data explained in this section.

The United Kingdom consists out of former sovereign states England, Wales, Scotland and Ireland. England is divided into 48 counties having an estimated population of about 56.0 million in 2020. In the south-east of England is the county Surrey with an area of 1,663 km^2 with a population of 1,189,934 (estimate in 2018, in our database 1.16M $\approx$ 1,161,256). This population can be found in about 29.1k postcode areas. This increases to 50.4k when buffer is included and all delivery points are considered. There are 500.6k delivery points (out of which are 481.4k domestic ones within the 29.1k postcode geometries). Including the buffer there are 800.8k delivery points (out of which are 760k domestic ones with 43.9k geometries). Delivery points, addresses and other general geographical data can be obtained from Royal Mail's Postcode Address File (PAF) and Ordnance Survey's portal. Please see BusAnalytics.uk for more details. The population used in the database is 1.16M and 1.86M including the buffer. The number of domestic delivery points which will be called households is 481.4k, i.e. there are about 2.41 people per household on average.

Our bus service information shows 37 bus operators (status 2018) in the county of Surrey. Several bus operators in Surrey participated in the subsequent case study. However, due to non-disclosure agreements (NDA) details about them cannot be mentioned. There are 270 routes with distinct bus service numbers operating in Surrey. Sometimes a bus service number is assigned to several slightly differing routes depending on the time or whether there were route deviations. In this study the most frequent (usually the longest) hat to be used. Furthermore, a subset of 87 routes (i.e. 32.2% of all routes) with distinct service number were analysed. Again due to the NDA these routes will not be identified. That means displayed routes to illustrate concepts are not necessarily related to data provided by bus operators. Information about bus stops and train stations can be obtained from the UK's government page by accessing the National public transport access node (NaPTAN) data (see BusAnalytics.uk).

Figure 1 gives an geographic overview of the area of interest. The heatmap was derived for the domestic households in Surrey (excluding buffer zone) using Kernel Density Estimators (KDE).

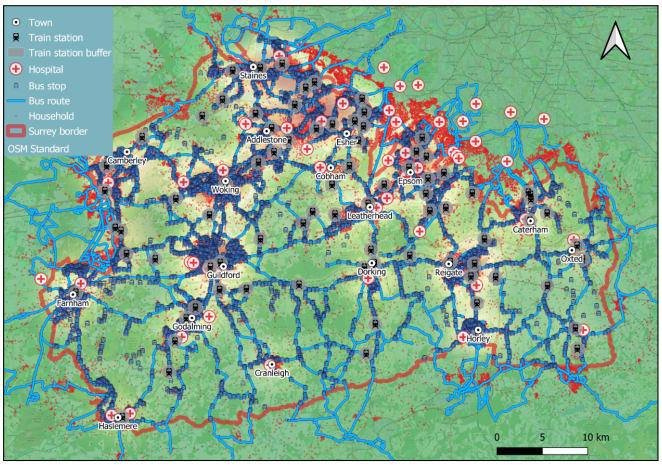

Figure 1: Surrey overview map.

⋆

Tel.: +44(0)1483 68 2005; fax: +44(0)1483 68 9511.

[1] It should be stressed out that this is very different to forecasting the number of passengers on existing routes. Forecasting the number of passengers on existing routes can be achieved using expert systems selecting the best method automatically. Popular forecasting techniques include MA, AR, ARMA, ARIMA and GARCH. The interested reader is referred to **?** or **?**.

1.1. Features

Bus operators, Trapeze and others provided data, which is stored in the BusAnalytics.uk database.

```
## Aggregated data R
##  [1] "bus_operator"   "passengers"     "route"          "weekday"
##  [5] "peak"           "sdate"          "deviation"      "headway"
##  [9] "households"     "hospitals"      "train_stations"

## Data for modelling
## [1] "passengers"     "weekday"        "peak"           "deviation"
## [5] "headway"        "households"     "hospitals"      "train_stations"
```

The above output shows the predictors (factors, features, variables) and *passengers* response variable. The aim will be to find a model that predicts the number of passengers based on the above features.

The provide data covers the period from 2015-11-16 until 2018-09-30 (i.e. about 2.92 years) and ~28M passenger journeys.

1.1.1. Passengers

The passenger numbers in data set R are aggregated values, which were derived from individual ticket data. The aggregation was achieved by grouping by route, weekday and peak time, compressing the data to 57,820 records. Weekday and peak time are defined below. Summary statistics of the aggregated data are shown below:

```
## Passengers grouped by route, peak/off-peak and weekday/end
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    49.0   174.0   485.6   632.0  8048.0

## Average number of passengers per day grouped by route
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.26   75.52  236.32  575.81  680.27 6550.10
```

75% of these aggregations have less than 632 passengers. The following profile (Figure 2) focuses on the daily routes with less than 1000 passengers. This represents 87.4% of all daily routes. 50% of all routes have less than

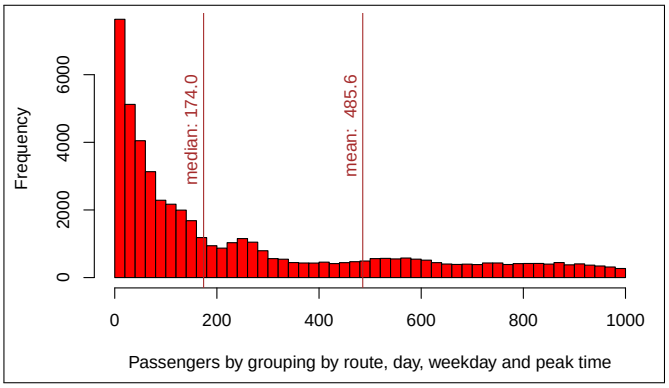

Figure 2: Passengers Histogram grouped by route, day, week-day/end and peak.

174 passengers on average during a day distinguished by peak and weekday (see below for details). In this discussion, a good practical guide to fitting probability distributions to existing data can be found.

It is also interesting to consider the average weekly number of passengers per route.

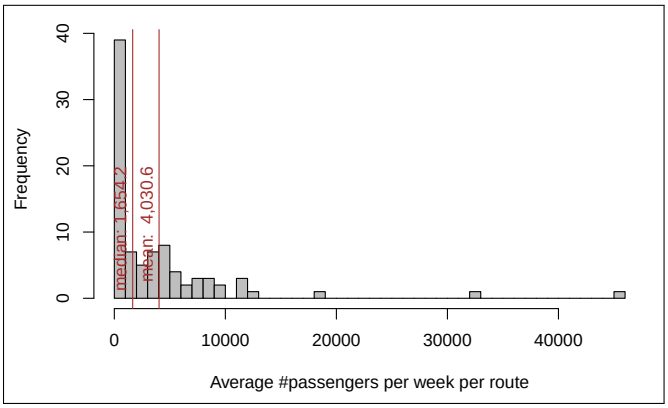

Figure 3: Histogram of average number of passengers per week.

As can be seen in Figure 3 there are three routes, which have significantly more passengers. However, 50% of the routes have less than 1,654.2 passengers per week on average.

1.1.2. Weekday

The *weekday* features distinguishes between two categories weekday and weekend. We investigated the differences between weekdays (Monday to Friday), but found that these are minor in comparison to weekends (Saturday and Sunday). Figure 4 displays the passenger volume distribution aggregated over the entire period.

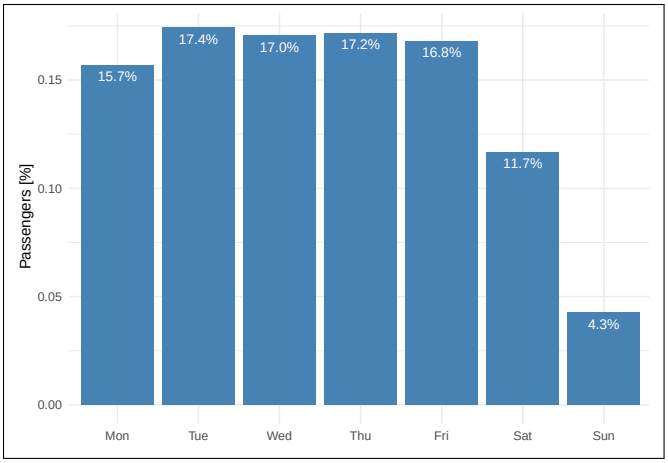

Figure 4: Weekdays volume.

A refined prediction model using each day (i.e. Monday to Sunday) as factor did not lead to significant differences. Hence, the simpler model was used. Note: a further im-

provement of the model can be achieved by classifying bank-holidays as weekends.

1.1.3. Peak

Peak and *off-peak* times are defined as 07:00-17:59 peak, and the remaining time as off-peak. The definition deviates from those given by bus operator, e.g. Transport for London: 6:30-9:30 and 16:00-19:00. The definition used here reflects the actual usage profile (Figure 5).

Figure 5: Passengers aggregated day profile.

It can be seen that around 7am (actually 7:10am) the average is exceeded, and at around 6pm (actually 18:17) the number of passengers is below the average. The peak at around 9:30am is explainable as this is the time, when off-peaks fare start usually. The peak at 15:30 could be related to passengers travelling home from work.

1.1.4. Headway

The predictor *headway* is the time difference in seconds between consecutive buses on the same route averaged over one day. The basic summary statistics are:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0167   7.5944  27.2131  38.3807  54.9419 585.7333
```

50% of daily peak/off-peak routes threshold is at 27.2 minutes (1,632.8s) headway. The average is 38.4min. Figure 6 shows an almost exponential distribution. The longest headway is 9.8 hours) during a day. The other extreme is a route about one second headway which could indicate that buses drive behind each other. Closer inspection of those cases is recommended.

1.1.5. Deviation

Deviation is the difference between actual and scheduled arrival of the bus at a stop. Here, the average deviation of the absolute differences for a day is used. This is measured in seconds.

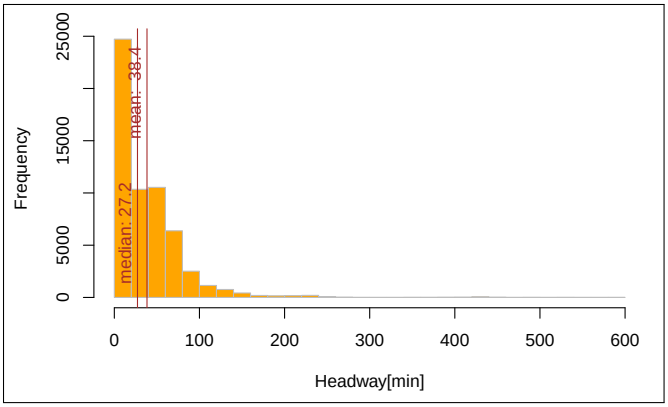

Figure 6: Histogram of headways.

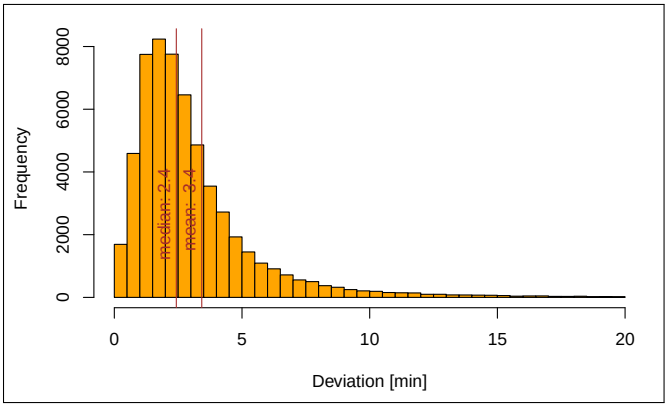

Figure 7: Histogram of deviations.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.527   2.428   3.424   3.763 208.602
```

```
##
##         Weekday Weekend
##    100        2       0
##    5001       2       1
##    6001      37      12
##    6005      13       6
##    6020       0       1
```

50% of all deviations are less than 2.4 minutes. The maximum deviation is 3.5 hours. Figure 7 shows the average deviations that are less than 3.4 minutes. This demonstrates that shorter deviations are more likely. The high deviations (≥ 1 hour) are limited to 80 routes and constitute only 0.1% of the total observations.

1.1.6. Fare

Fare is the average of all above-zero payments within a day on a route measured in pounds. That means, there are passengers who do not have to pay. It includes payments for weekly, monthly or season tickets. These payments distort the daily fare scale. In theory those should be filtered or adapted according to their valid period. In practice this is a difficult undertaking. However, two potential improvements can be done (1) to create a fare look-up table to allow adaptations; (2) to filter out payments which exceed single tickets. Furthermore, fares are different for different travel zones along a route. This should be well covered by taking the average. Overall, this factor is - at the moment - problematic and it is expected that machine learning techniques will filter it out.

1.1.7. Hospitals and train stations

Hospitals are the number of hospitals within an 800 meters beeline-distance to the route. In this study the hospitals within Surrey county and the ones close to its border were used.

Train-stations are the number of train stations within 800 meters beeline distance to the route.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.912   4.000  12.000
```

Detailed visualisation of the factors can be found in the `Stageocoach-model-data.pbi` located in the `PBI` folder. Correlations and detailed analyses will be available in `cor.R`

Table 1: Hospitals and train stations per route.

route	hospitals	train stations
21	6	9
420	6	12
460	5	9
134	4	5
135	4	6
26	4	2

Table 2: Sample of *R*.

	passengers	deviation	headway	households	hospitals	train_stations
16079	819.0	2.2	7.8	20,517.0	1.0	5.0
55517	1,003.0	2.1	15.3	12,274.0	3.0	1.0
44255	180.0	3.6	51.7	11,756.0	1.0	2.0
196	20.0	2.1	50.9	15,537.0	2.0	2.0
42148	123.0	3.4	54.6	19,050.0	2.0	3.0
35223	51.0	8.5	112.7	13,143.0	0.0	6.0

1.2. Correlations

Before we create the correlation matrix a brief data extract is shown in the table below.

The correlation matrix states that headways have an observable correlation to passenger of -25.8%. That means, shorter the headways the more passengers, which makes perfect sense. Train stations and passengers have correlation of 15.6%, which indicates that bus passengers need to use trains for their journeys. All other features have surprisingly low correlation to passengers. Housholds are highly correlated to hospitals and train stations, which indicates that routes passing through populated areas (high number of households), are also passing by hospitals and train stations.

As an alternative approach all route data was aggregated by bus operator, service (route), peak and weekday. The average daily data was averaged again and the standard deviation of the daily averages was determined for passengers, headways and deviations. The households, hospitals, train stations and route length had no variations.

Figure 9 shows the correlations for all distinct routes for weekdays at peak time. In total there were 86 observations in this data set. The passengers mean, median and

Table 3: Correlation matrix.

id	passengers	deviation	headway	households	hospitals	train stations
passengers	100.0%	-0.4%	-25.8%	7.1%	-0.6%	15.6%
deviation	-0.4%	100.0%	6.7%	-1.6%	-1.9%	-2.7%
headway	-25.8%	6.7%	100.0%	-24.3%	-18.6%	-29.4%
households	7.1%	-1.6%	-24.3%	100.0%	78.5%	86.2%
hospitals	-0.6%	-1.9%	-18.6%	78.5%	100.0%	74.4%
train_stations	15.6%	-2.7%	-29.4%	86.2%	74.4%	100.0%

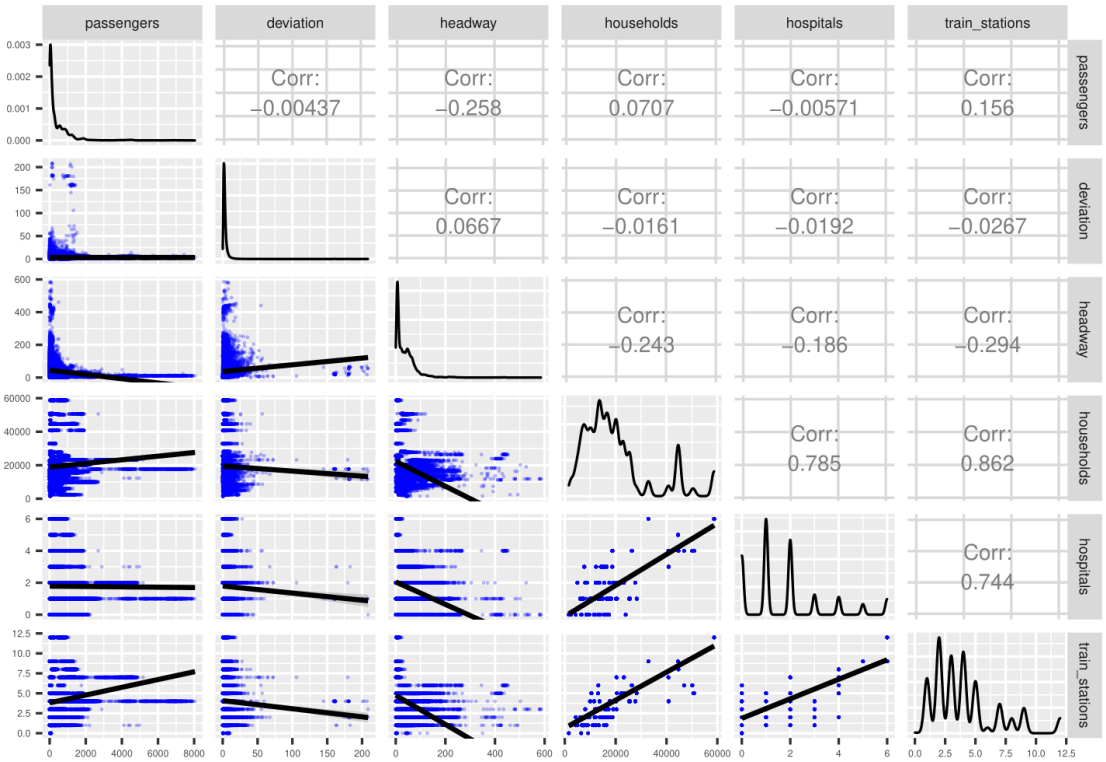

Figure 8: Correlation matrix details .

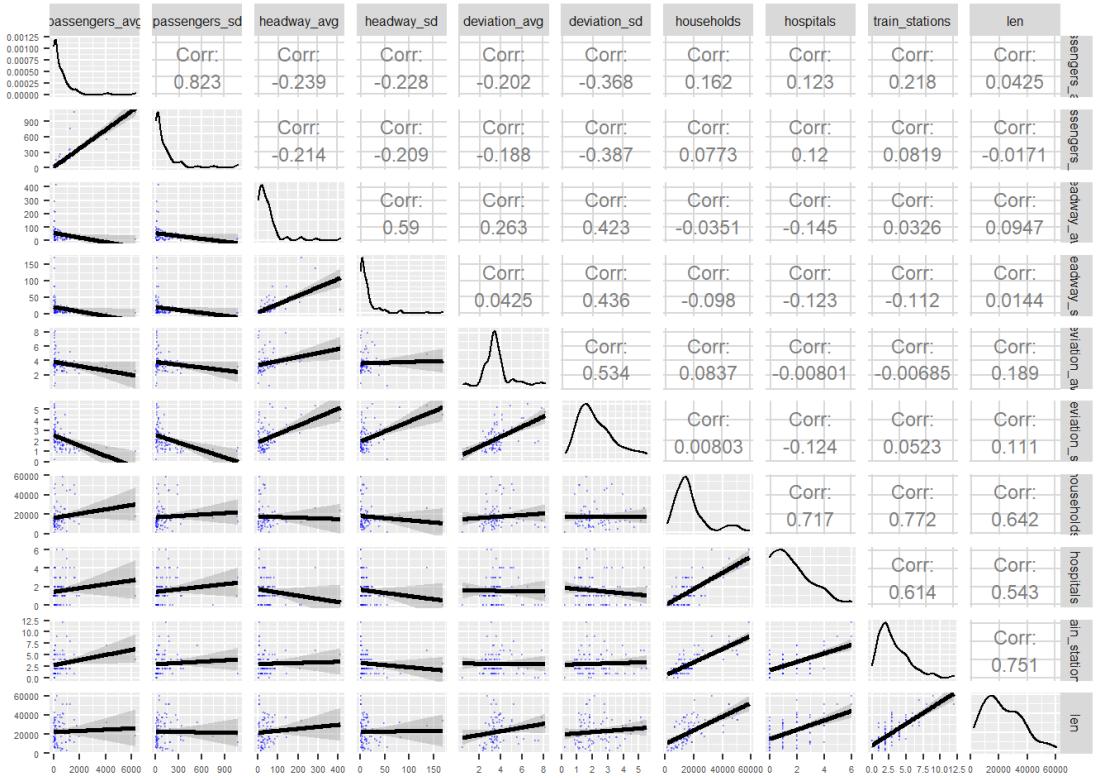

Figure 9: Correlations for aggregated route data..

maximum are 552.6, 251.9 and 6,357.5 respectively with a mean (median) standard deviation of the daily passenger averages of 121.0 (45.1). As expected all standard deviations show high correlations to their respective means.