Chapter 5: Results & Discussion

Draft version

This chapter presents and discusses the results of the experiment described in Chapter 4. It is structured as follows. The first section shows the clusters that resulted from the cluster loop, along with their main characteristics, and the chosen locations of the model points. Section two provides the structure of the models that were build in the model loop, and the residual diagnostics for each them. Then, the third section covers the accuracy of the forecasts, while in the fourth and last section, the limitations of DBAFS are discussed, and recommendations for possible improvements are given.

5.1 Clustering

Figure 5.1a shows the grid overlaying the JUMP Bikes system area in San Francisco, including the centroids of all the grid cells. In total, the grid contains 249 grid cells, each 500 meter high and 500 meter wide. Figure 5.1b shows the calculated number of pick-ups per grid cell, during the training period. On average, there were 938 pick-ups per grid cell. The maximum number of pick-ups in a grid cell was 8189, while in 14 of the 249 grid cells, there were no pick-ups at all. It can be seen that high counts of pick-ups occur in the grid cells along the diagonal axis from south-west to north-east. Mainly in the south-eastern corner of the system area, the usage intensity is very low.

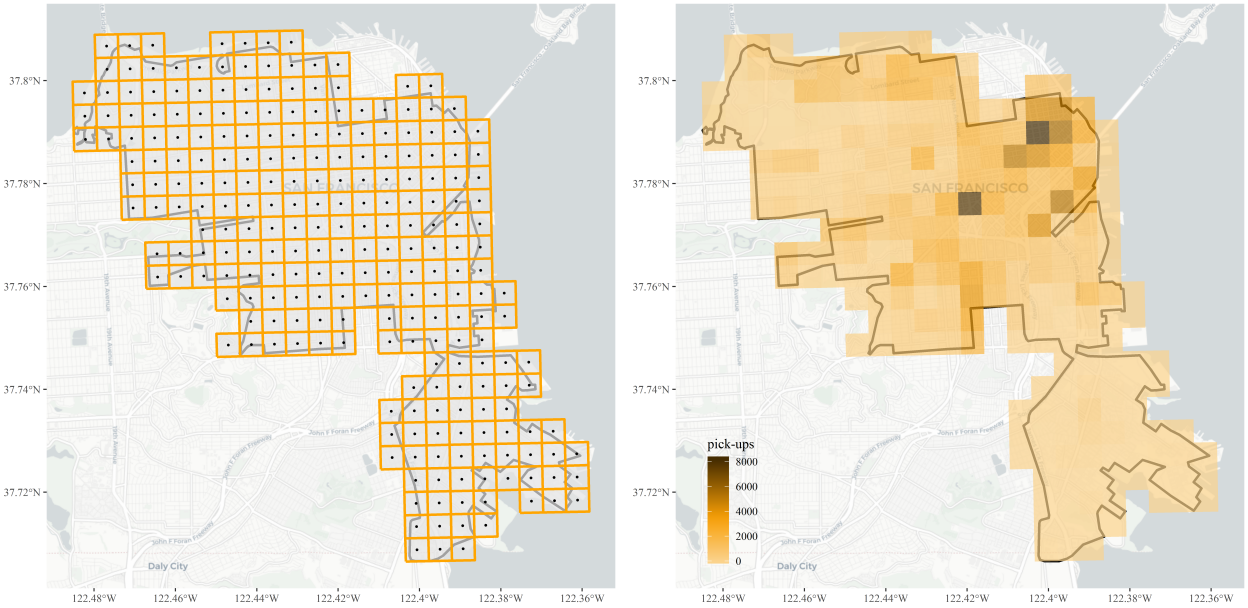

Figure 1: a) grid overlaying the system area; b) number of pick-ups per grid cell

Recall that for each grid cell centroid, a time series of historical distance data was queried, and that the average weekly, normalized patterns in these data were clustered using spatially constrained hierarchical clustering. The automatic procedure of defining the number of clusters k and the mixing parameter α, lead to a definition of $k = 4$ and $\alpha = 0.6$. This resulted in a partition containing four fully spatial contiguous clusters. The geographical outlines of these clusters are shown in Figure 5.2a. The centroid of all grid cell

centroids in each cluster, weighted by the number of pick-ups in the corresponding grid cells, are shown in Figure 5.2b. These weighted centroids serve as the model points in DBAFS.

Roughly speaking, and based on a large study of neighborhood indicators in San Francisco (San Francisco Department of Public Health 2014), the four clusters can be characterized as follows. The orange cluster covers the Bayview/Hunters Point neighborhood, which is a rather isolated area, with a high percentage of low-income households and relatively high crime rates. The blue cluster forms the city center of San Francisco, containing the neighborhoods with the highest population densities, but also with a relatively high job density compared to the residential density, and large areas zoned for commercial usage. The purple cluster mainly contains neigborhoods where the residential density is high compared to the job density, and the area zoned for commercial usage is relatively small. Finally, the green cluster covers the Presidio Park, a recreational area with few inhabitants, and a relatively high number of bike lanes. For the sake of clarity, the orange, blue, purple and green clusters are from now on refered to as the *Bayview*, *Downtown*, *Residential* and *Presidio* clusters, respectively. Consistently, the four corresponding model points will be called the *Bayview*, *Downtown*, *Redisential* and *Presidio* model points, respectively.

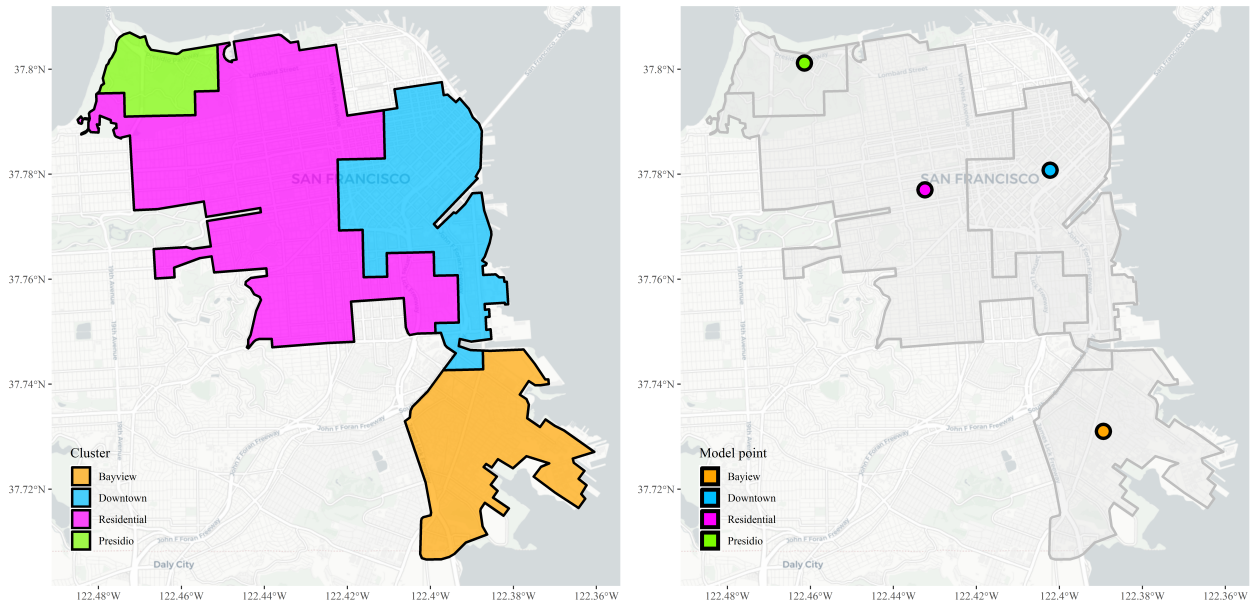

Figure 2: a) geographical outline of the clusters; b) geographical locations of the model points

Table 1 presents some descriptive statistics of the time series, averaged per cluster, and averaged over the whole system area. From the 249 grid cells, almost half are located within the Residential cluster. The Bayview and Downtown cluster are both medium sized, while the Presidio cluster is by far the smallest of the four.

During the training period, the nearest available bike was on average located 529 meters from the grid cell centroids. In the Bayview cluster, this distance was more than twice as high as in the other clusters, that did not vary much between each other. The Bayview cluster also showed the largest variation in the data, with a high average standard deviation compared to the other clusters, and an average range that spans more than four kilometers. This can possibly be explained by the low usage intensity of the bike sharing system in this part of the system area. When the number of bikes in an area is low, the nearest available bike and the second nearest available bike are more likely to be far away from each other. In that case, when the closest of them gets picked-up, the distance to the nearest available bike will suddenly increase substantially. The other way around, when all available bikes are far away, and one bike gets dropped-off inside the area, the distance to the nearest available bike will suddenly decrease substantially.

Although not as extreme as the Bayview cluster, also the other clusters had on average high ranges when compared to the mean and standard deviation. However, the standard deviation itself turned out to be rather small relative to the mean. This implies either the presence of outliers, or population distributions with thin, but wide tails.

The first order autocorrelation measures the average dependency between data values at time t and correpsonding data values at time $t-1$. In the whole system area, this dependency was strong, especially in the Bayview and Presidio clusters. These high autocorrelation values are important, since they imply that the data are not random, and it is reasonable to use past observations when forecasting future ones. However, the calculated spectral entropy values show that in general, the data are also very complex, and the forecastability is low. This mainly concerns the Downtown and Residential clusters, which contain, as could be seen in Figure 5.1b, the areas where the pick-up density is high. In such areas, the data are more dynamic, since bikes get picked-up and dropped off constantly, and the location of the nearest available bike will change often. In most cases, the more dynamic the data, the harder to forecast.

Table 1: Descriptive statistics of the grid cell centroids distance data

	N	μ	$range$	σ	$\rho(1)$	H
Total	249	529	2828	155	0.84	0.77
Bayview	54	903	4152	169	0.96	0.66
Downtown	58	451	2521	62	0.78	0.82
Residential	120	415	2512	88	0.81	0.81
Presido	17	410	1905	85	0.91	0.72

Note:
Except N, all metrics are calculated for each time series seperately, and averaged afterwards.
[1] N is the total number of grid cell centroids
[2] μ is the mean of the data, in meters
[3] $range$ is the difference between the maximum and minimum data value, in meters
[4] σ is the standard deviation of the data, in meters
[5] $\rho(1)$ is the first order autocorrelation, see section 2.x
[6] H is the spectral entropy, see section 2.x

Figure 5.3 shows the normalized, average weekly patterns of the time series, averaged once again per cluster. The patterns can be explained intuitively. The Bayview cluster has a low usage intensity, and although there are peaks in the data every day, a clear and consistent pattern is absent.

The Downtown cluster has a high density of jobs and commercial activities. During working hours, the demand for bikes is low, which leads to a high number of available bikes, and consequently, short distances to the nearest available bike. In the afternoon, just after working hours, the demand starts increasing, and it gets harder to find an available bike nearby. Later in the evening, the demand decreases again. In the weekends, the daily peaks are less clear, an spread out over a longer timespan.

The Residential cluster shows the exact opposite pattern. In the morning rush hours, commuters use the bike to get to work, and not many available bikes are left in the residential areas. Hence, in those areas, the distance to the nearest available bike is higher during working hours. In the afternoon, commuters come back from work, and leave the bikes in the residential areas, causing a decrease in distance to the nearest available bike. Just as in the Downtown cluster, the peaks and valleys in the weekends are less strong. Furhtermore, they happen later on the day, corresponding to the same periods as the Downtown cluster.

Finally, the Presidio cluster is mainly a recreational area. There are a lot of bikes, but during weekdays, they are used less, leading to small and constant distances to the nearest available bike. In weekends, and mainly on Sunday afternoon, the usage intensity is high, and it takes longer to find an available bike.

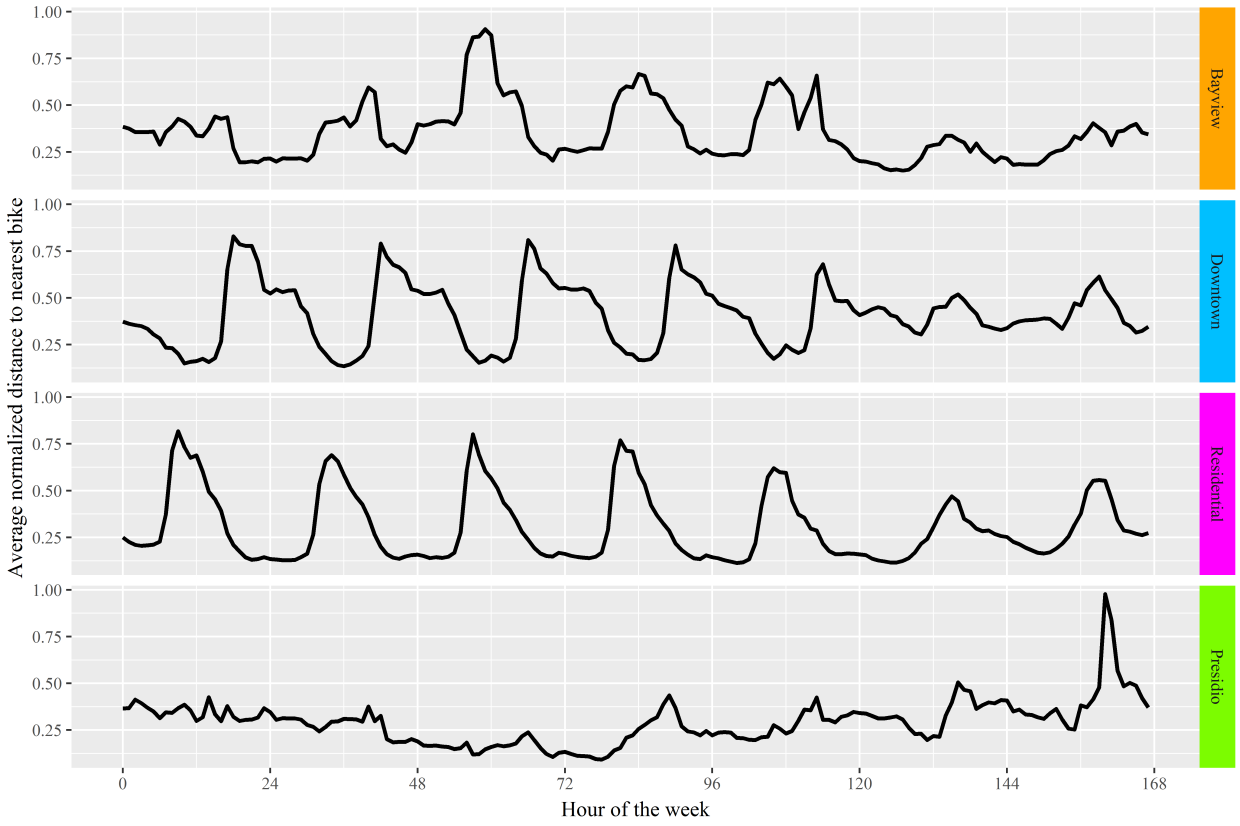

Figure 3: Normalized, average weekly patterns of the grid cell distance data, averaged per cluster

5.2 Model building

Figure 5.4 shows the time plots of the distance data that were queried for each of the model points in Figure 5.2b, with the darkgrey shaded areas representing weekends. The plots endorse the findings in the previous sections. The data corresponding to the Bayview modelpoint show large variation, interspersed with flat sections, and lack a clear repeating pattern. The variance decreases to the end of the training period, together with the mean.

The data corresponding to the Downtown and Residential modelpoints are most dynamic. A daily pattern shows for both of them, and in the Downtown data, the peaks in the weekend are generally lower than those on weekdays. For the Residential modelpoint, however, this difference is less visible. In both datasets, the dialy peaks vary considerably in height from day to day, and look far less smooth than the averaged weekly patterns shown in Figure 5.3. Although not as noticable as for the Bayview modelpoint, the variances of the Downtown and Residential data seem to decrease slightly towards the end of the training period, together with the mean. This justifies the use of a multiplicative decomposition, rather than an additive one.

The Presidio modelpoint shows the most constant data, with a low mean and long flat sections. In the first part of the training period, the Sundays stand out clearly, but later, the peaks in the data get smaller, and seem to occur more randomly.

Finally, what strikes, is the large period of missing data in the last week of the training period. These data are missing in all datasets, probably caused by problems on the JUMP Bikes server. Some smaller periods of missing data occur in the first weekend of the training period, and during the second week of November. Again, the data for all four time series are missing in these periods. Recall here that both STL

and the implementation of ARIMA in R, as used by DBAFS, automatically handle missing values. Hence, the periods of missing data are not problematic for neither the model fitting nor the forecasting process.

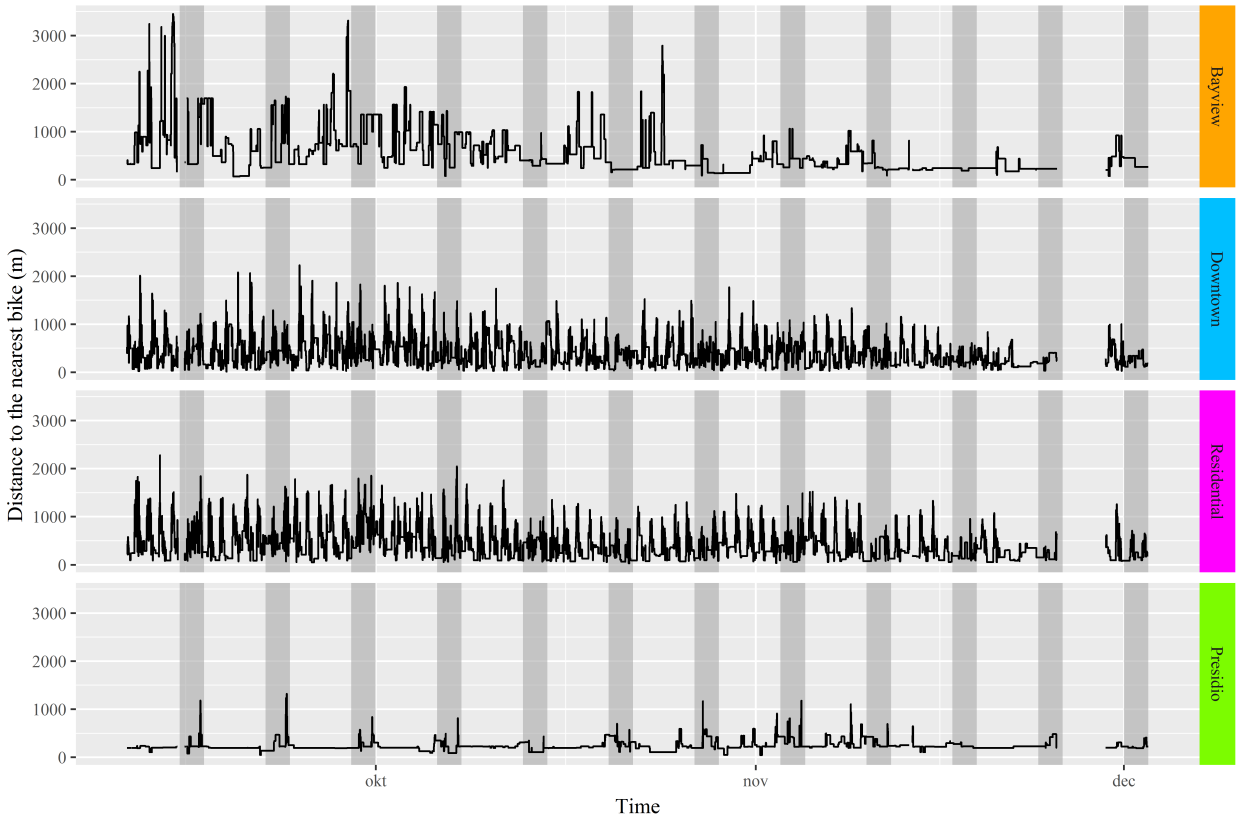

Figure 4: Time plots of the model points distance data

The structures of the models that were fitted on the four model point time series are shown in Table 5.2. The automatic seasonality detection resulted, as expected, in a double seasonal pattern for the Downtown data. For the Residential data, only a daily pattern was detected, while also the Bayview data were considered seasonal by the cross-validation process. Only Presidio data were labeled non-seasonal, and skipped the decomposition process sequence. Additionaly to the seasonal periods, Table 5.2 also shows the strength F_s of the corresponding seasonal patterns, calculated with Equation 2.x. The daily pattern in the Bayview data is weak, just as the weekly pattern in the Downtown data. The seasonal strengths of the daily patterns in the Downtown data and the Residential data are larger, but still, none of them can be considered very strong.

In the ARIMA(p, d, q) models for the Bayview and Downtown data, no autoregressive terms are included. Instead, they contain a high order of moving average terms. All datasets passed the KPSS test for stationarity after one differencing operation. The full details of the components and fitted models, including parameter estimates, can be found in Appendix C.

Figure 5.5 shows the residuals of each model, plotted over time. All models have residuals with an approximately zero mean, and the variances look approximately constant. Comparing the residual time plot with the data time plot of Figure 5.4, it can be seen that in for the less dynamic data of the Bayview and Presidio model points, the models struggle to find a good fit for the peaks and valleys in the data, while the flat sections are explained accurately.

Table 2: Model structures

	STL		ARIMA		
	seasonality	F_s	p	d	q
Bayview	daily	0.25	0	1	5
Downtown	daily and weekly	0.40, 0.27	0	1	5
Residential	daily	0.47	2	1	1
Presidio	none	-	1	1	1

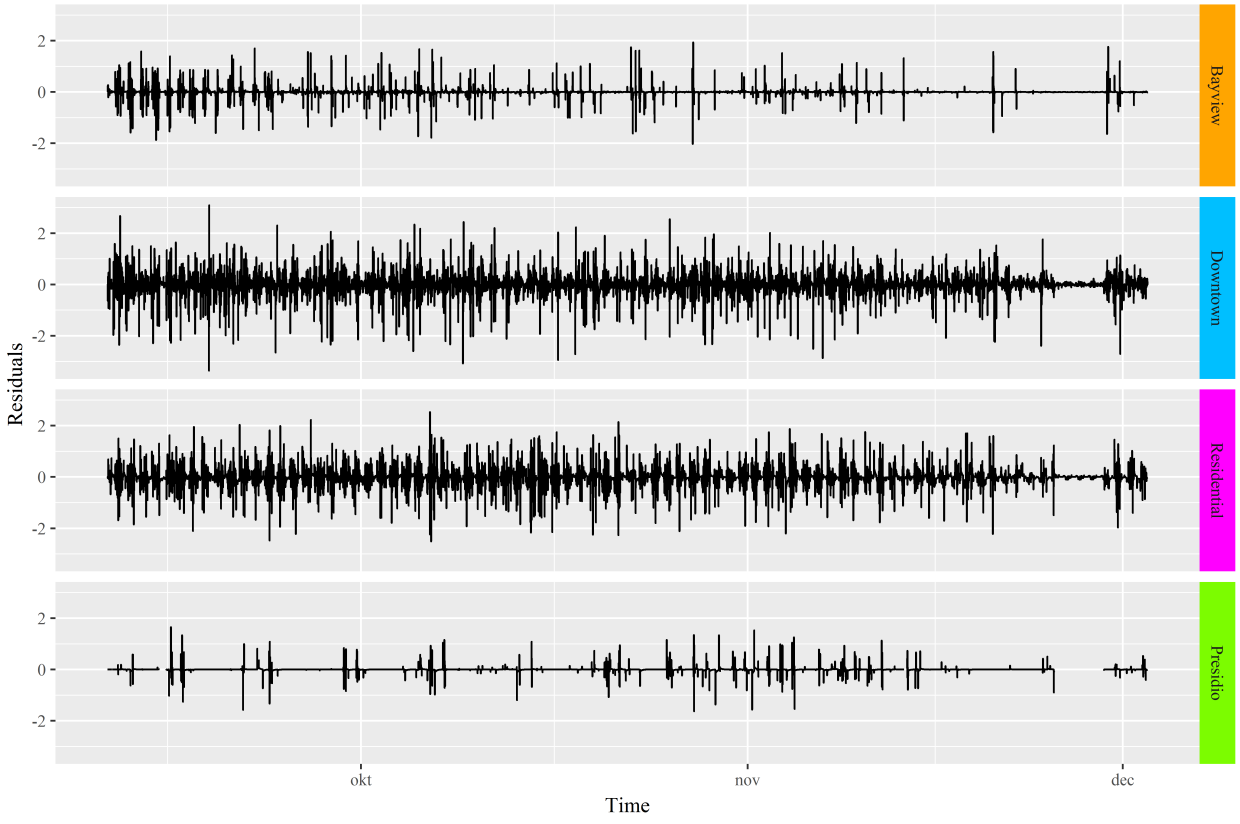

Figure 5: Time plots of the model residuals

The autocorrelations at several time lags in the residuals are shown in Figure 5.6. Since the data have a temporal resolution of 15 minutes, 96 time lags correspond to one day, and 672 time lags, the total span of the x-axis in the figure, to one week. The dotted orange lines form the lower and upper 95% confidence bounds, assuming a normal distribution. This means that the residuals are considered to be a realization of a white noise process when at least 95% of the autocorrelation values fall within these bounds. It is important to note here that when working with real-world data, finding perfectly random model residuals is an exception, especially when the data have a high entropy. Taking that into account, the autocorrelation plot of the Downtown model look good, and their residuals seem to approximate white noise.

However, for the Bayview and Residential models, and in lower extent also the Presidio model, autocorrelations show peaks at several lags that correspond to full days. This implies that there is still some information left in those data that the models did not capture. Recall that no weekly seasonal pattern was included in the Bayview and Residential models. Only using a daily seasonality, systematic differences between weekdays

will not be explained, which could be a possible reason for the significant autocorrelations that show up. In the Presidio, no seasonal pattern was identified at all. However, the seasonal patterns were chosen such that the RMSE of the forecasts in the cross-validation process was minimized, and in the end, minimizing these forecast errors is the real aim of DBAFS.

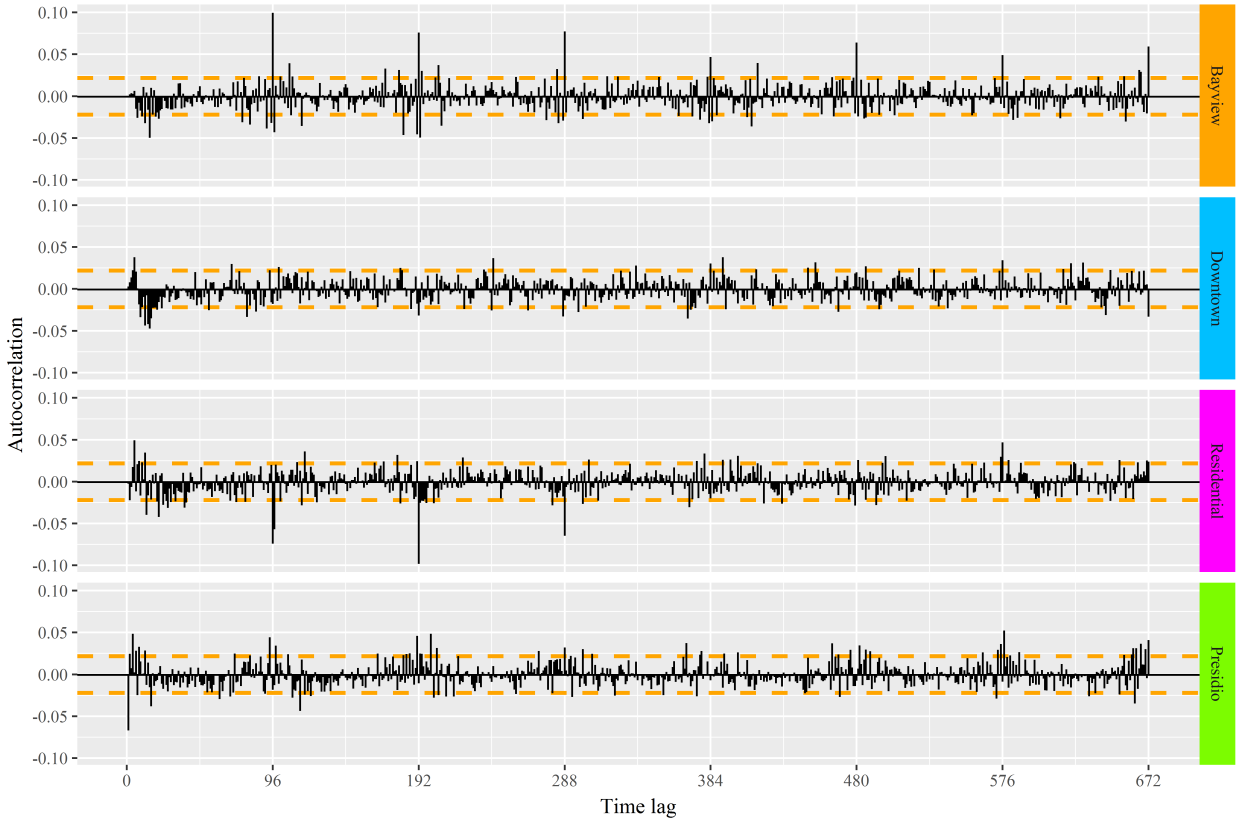

Figure 6: ACF plot of the model residuals

Finally, Figure 5.6 shows the histograms of the model residual distributions. As expected, for the Bayview and Presidio models, most values are clustered closely around the zero mean, with the tails being extremely thin and long, especially for the Bayview model. The residuals of the Downtown and Residential models follow a distribution that comes closer to a normal one, but also here, the tails are wide.

As discussed in section 2.4.2.4, using Gaussian likelihood is sensible even when non-normally distributed residuals show up. Of course, it could be a possibility to try different likelihood functions, but this will make the process much more complex and much slower, for, probably, only a little gain in forecast accuracy. The non-normality of the residuals does have an effect on the validity of the prediction intervals, however. When 95% prediction intervals are calculated, assuming normality of the forecast distribution, they can not be interpreted as such. This issue will be covered further in the next section.

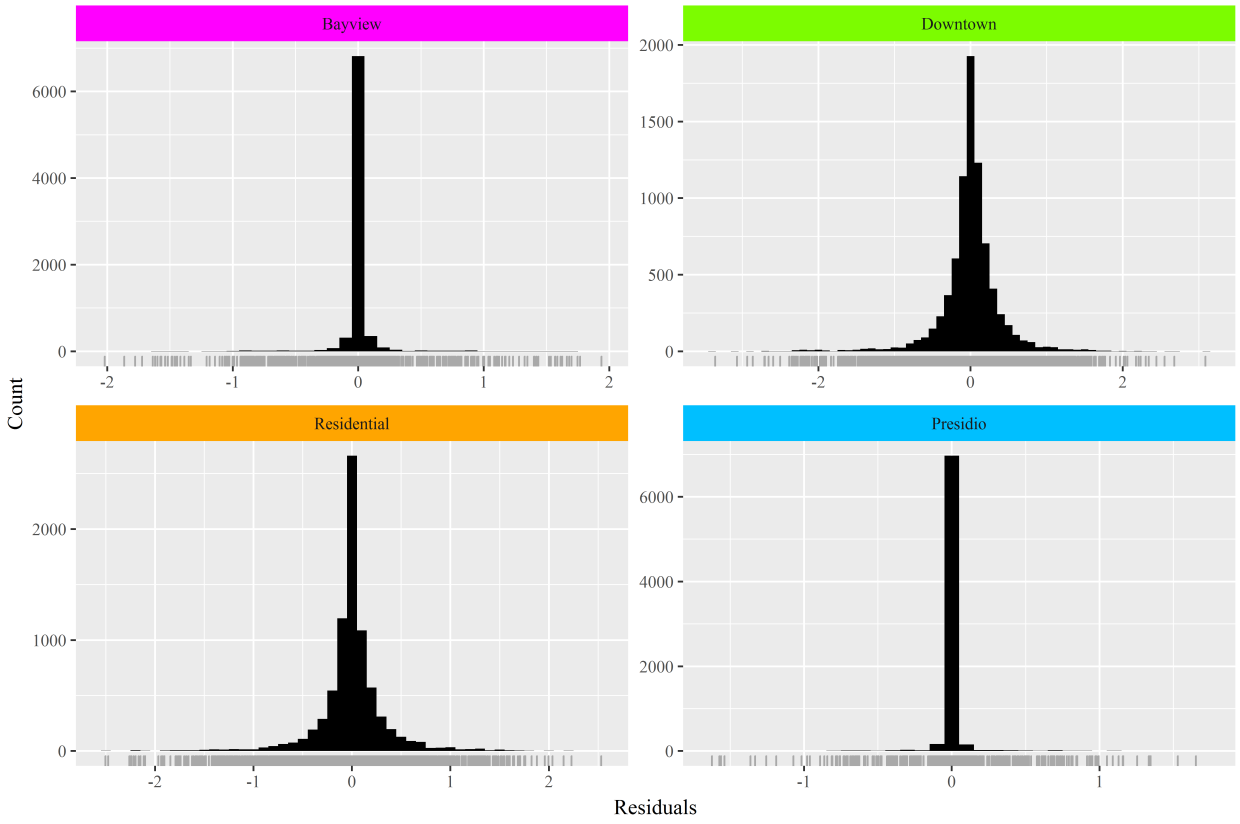

Figure 7: Histograms of the model residuals

5.3 Forecasting

Figure 5.6a shows the geographical locations of the test points. The full information on the test set, with all unique location-timestamp combinations, can be found in Appendix D. Table 5.3 lists the average RMSE of the forecasts, both of DBAFS and the naïve method, for respectively the total system area, and per cluster.

5.4 Limitations & Recommendations

References

San Francisco Department of Public Health. 2014. "The San Francisco Indicator Project." https://www.sfindicatorproject.org/.