Improving French railway service

N. Francescon, A. E. Franzoni, E. Garlanda

16th February 2024

Contents

Introduction

1.1 Goal of the project

The aim of the project is to support the decisions of SNCF (Société Nationale des Chemins de fer Français), which is the company that manages the French railway system. The company provides 5462 observations [1] about TGV trains over a period of 47 months (January 2015 - November 2018) for 130 routes. SNCF needs to understand which interventions would improve the service in terms of delays and cancellations in 2019.

1.2 Dataset description

The data were published by SNCF at the end of 2018. The dataset includes the following variables:

- year and month
- departure and arrival station
- type of service (national/international)
- total number of trips
- average journey time
- total number of cancellations
- total number of late departing trains and their average delay at departure
- average delay at departure of all trains
- total number of late arriving trains and their average delay at arrival
- average delay at arrival of all trains
- causes of the delay at arrival
  - external causes
  - rail infrastructure causes
  - traffic management causes
  - rolling stock causes
  - station management causes
  - travelers causes
- number of trains with more than 15, 30 and 60 minutes of delay at arrival
- average delay at arrival of the trains with more than 15 minutes of delay

Each row of the dataset is identified by the departure station, the arrival station and the month-year pair. The data are already aggregated: there is no information about single trips. Each row of the dataset is related to the trains of a route in a specific month. Moreover, all the collected data are related to TGV (Train à Grande Vitesse) trains, i.e. high-speed trains.

[1] The dataset can be found at: https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-02-26

1.3 Data preprocessing

For convenience, a column indicating the route, i.e. the pair departure-arrival station is added.

Looking at the data, some rows show very unlikely observations. In particular, some routes have a negative average delay at arrival. Such a behavior is reasonable if the train arrives on average a few minutes early. However, 7 recordings report an early arrival of more than 30 minutes on average. This is quite unrealistic, especially keeping in mind that data are aggregated. Therefore, such observations are removed from the dataset.

In order to develop the analysis, further data aggregation is performed. Data are aggregated by year, so that each row of the dataset corresponds to a route in a given year. This approach is justified by the fact that the company is mainly interested in the overall performance over the 4 years, without a specific focus on the month. Moreover, the original dataset was already aggregated, so the loss of information caused by a further aggregation step is limited.
To merge the original observations, suitable summations and weighted averages are performed, according to the nature of the covariate.
In the original dataset, routes are observed over 47 months. However, 22 routes have more than 24 missing months of observations. These are removed in building the aggregated dataset, in order to keep the analysis consistent.
A final dataset composed of 108 routes along 4 years is obtained.

Exploratory analysis

Once the dataset is built, some preliminary analysis is conducted.
The focus is on the average delay at arrival, which at a first glance appears to be a significant indicator of performance for the company.

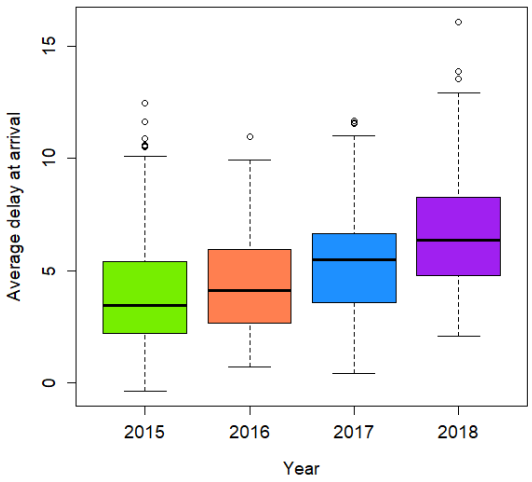


Figure 2.1: Boxplot for the average delay at arrival

The Figure 2.1 suggests an increasing trend in the average delay at arrival. To further investigate this aspect, an ANOVA test is performed to compare the mean average delay at arrival along the 4 years.

$$H_0 : \mu_{2015} = \mu_{2016} = \mu_{2017} = \mu_{2018} \qquad H_1 : \exists i, j \in \{2015, 2016, 2017, 2018\} : \mu_i \neq \mu_j$$

Since data are not normal within each group, the assumptions of parametric ANOVA are not met. Thus, a permutational approach is employed, using the F-statistic. The resulting p-value is 0, so the null hypothesis is rejected for any reasonable level of significance: there is a significant difference in the means over the years.
To better understand how the has delay evolved over the years, Bootstrap t-intervals ($\alpha = 0.05$) for the mean average delay at arrival are built applying a Bonferroni correction (k=4).

	lower	center	upper
2015	3.73	4.37	5.24
2016	3.89	4.47	5.07
2017	4.97	5.48	6.09
2018	6.16	6.84	7.56

The increasing trend is confirmed by the confidence intervals: the lower bound in 2018 is greater than the upper bounds of all the previous years. Therefore, SNCF should be aware of this loss of performance.

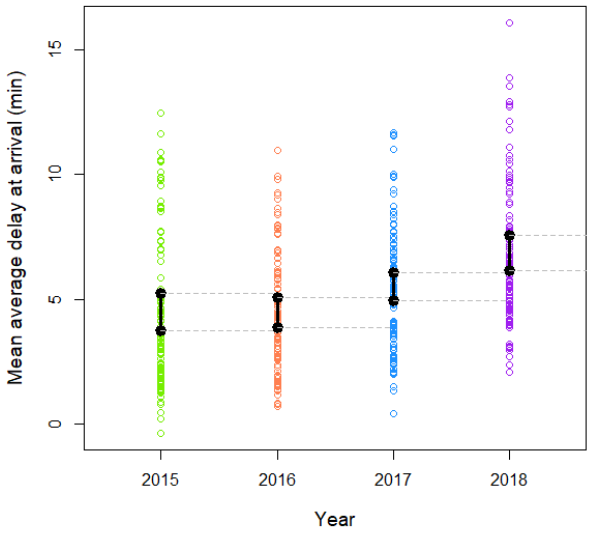
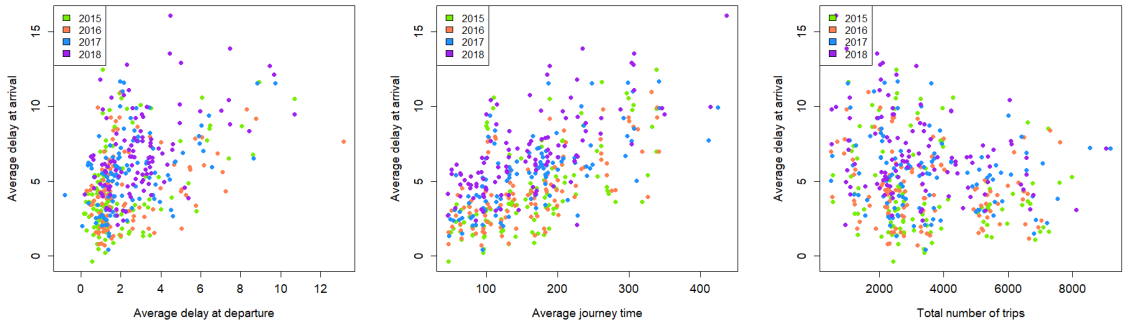


Figure 2.2: 95% Bootstrap t-intervals for the mean average delay at arrival

Explaining delays

After evidencing the increasing trend, a tentative regression model is built. The dataset presents a few variables that may provide a better explanation of the average delay at arrival. In particular, the average delay at departure, the total number of trips and the average journey time are considered. Other variables are excluded because of redundancy of information - that could lead to high correlation among the covariates - or because they do not seem to carry relevant information about the delays. Moreover, a strong consideration needs to be stated: data originated from the same route along the years cannot be considered independent. Thus, the regression model must include the year of observation as a covariate both in terms of group-depending intercept and of interaction for all the other covariates to preserve the independence assumption.



(a) Scatterplot of average delay at departure vs at arrival

(b) Scatterplot of average journey time vs average delay at arrival

(c) Scatterplot of total number of trips vs average delay at arrival

From the plots there seems to be a relation between the average delay at arrival and both the average delay at departure and the average journey time. No pattern is noticeable considering the total number of trips, therefore the variable is not included in the regression model.
A first simple parametric model is fitted:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \beta_{2j} z_{ij} + \epsilon_{ij} \qquad i = 1, ..., 108 \quad j \in \{2015, 2016, 2017, 2018\}$$

where y_{ij} represents the average delay at arrival for the i-th route in the j-th year, x_{ij} denotes the average delay at departure and z_{ij} is the average journey time.
Although the R^2 of the fit is decent (0.68), the model shows some weaknesses: the residuals in Figure 3.2 appear neither homoscedastic nor gaussian distributed, but above all there is high collinearity between the considered variables. A more simplified version of the model would lose the independence assumption and cannot be fitted.

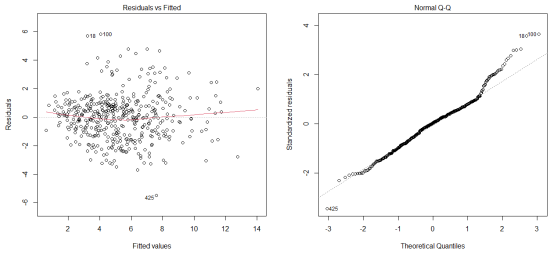
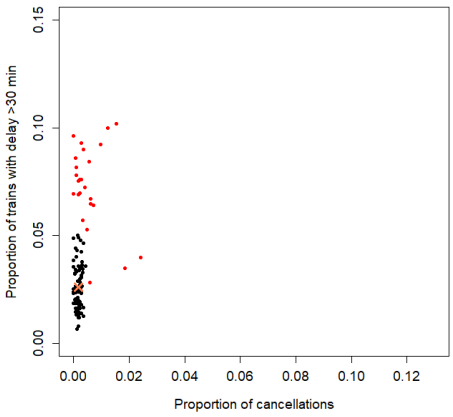


Figure 3.2: Diagnostics of residuals

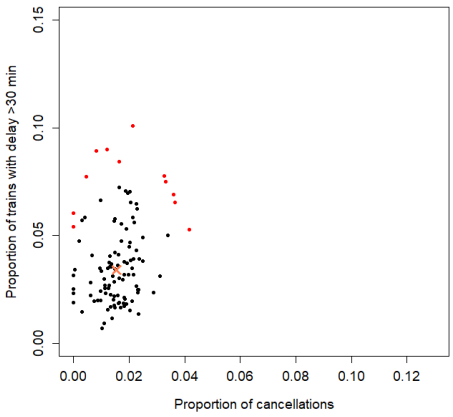
The main application of the model would be to predict the average delay at arrival given the average journey time and the average delay at departure. For instance, this information could be included in the app of the company to let travelers know how much delay to expect. However, this is not possible since the year is included as a covariate and also as interaction term to keep the independence assumption, making the model unable to predict anything related to future years.

Cancellations and extreme delays

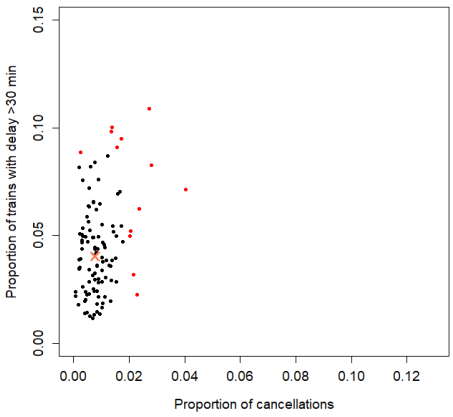
From now on, the interest moves towards outlier detection since regression models do not give satisfactory results. In addition, outlying routes should be detected to be able to report SNCF which are the most critical aspects of its railway network.
An interesting aspect for the analysis involves the number of canceled trains and the number of trains with a delay greater than 30 minutes. Indeed, in these two cases, SNCF must give a partial refund to the travelers [1], [2], leading to an economic loss. The quantities are considered in percentage with respect to the total number of trips for each route and year.
To detect outliers, a robust analysis for bivariate data is developed. In addition, this method allows to compute the robust estimates of locations.
The four years are considered separately, to keep observations independent. For each year, the Minimum Covariance Determinant estimator with a reweighting step is computed via FAST-MCD algorithm, setting $\alpha = 0.75$ and using 2000 subsets for initial estimates. As a result of this procedure, a robust location estimate and some outlying routes are obtained for each year. The goal is to compare outliers over the years.



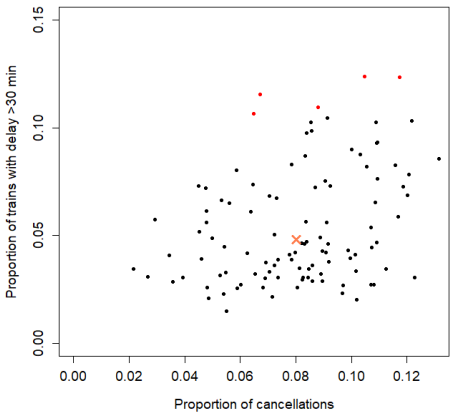
(a) Multivariate robust analysis in 2015



(b) Multivariate robust analysis in 2016



(c) Multivariate robust analysis in 2017



(d) Multivariate robust analysis in 2018

Figure 4.1: Outliers after the reweighting step are marked in red while the robust estimated location is marked with a cross

As a preliminary result of the four equiscaled plots, it can be noticed that no outliers with a positive behavior are detected: only outliers that show a high proportion of cancellations or extreme delays are recorded. After having noticed a wide increase in the proportion of cancellations, a further investigation over the robust location estimates is summarized by the following table.

	% of canceled trains	% of delays >30 minutes
2015	0.2%	2.6%
2016	1.6%	3.4%
2017	0.8%	4.0%
2018	8.0%	4.8%

Table 4.1: Robust estimates of location

The extreme delays appear increasing along the years, but the most critical aspect appears to be that in 2018 the percentage of canceled trains is much higher than in the previous years. Therefore, further information was searched to try to understand whether some external factors may have played a role in such a high cancellation rate.

Dealing with strikes

From external knowledge [3], in 2018 there were some strikes following a government proposal to liberalize SNCF. Such information is not included in the dataset and possibly other strikes happened over the four years. However, such events may have a detrimental effect on the analysis, because the company cannot act upon this type of cancellations and they may mask other problems in the service.

Strikes are generally announced for a specific day or even for some hours of a day, therefore ideally daily data about cancellations should be used. Since such data are not available, the best way to detect high cancellation rates - that may be related to strikes - is to consider the original dataset, which is aggregated by month.

To isolate months with a higher cancellation rate, Bootstrap reverse percentile confidence intervals ($\alpha = 0.05$) for the median cancellation percentage are computed. A Bonferroni correction is applied, with k=47, which is the number of observed months.
A month is considered critical if the lower bound of the related confidence interval is greater than the upper bound of the previous or the following one. In particular, the following months are detected as critical:

- March 2016 [4]
- April 2016 [5]
- June 2016 [6]
- December 2016
- January 2017
- March 2018 [7]
- April 2018 [3]
- May 2018 [8]
- June 2018 [9]
- July 2018 [10]

The months in which the sources report news about strikes are then removed from the dataset. Instead, the other months are kept because the high rate of cancellations may depend on the service of the company.

A new dataset is built, by removing the months with strikes and aggregating by year, as done at the beginning of the analysis.

Final analysis

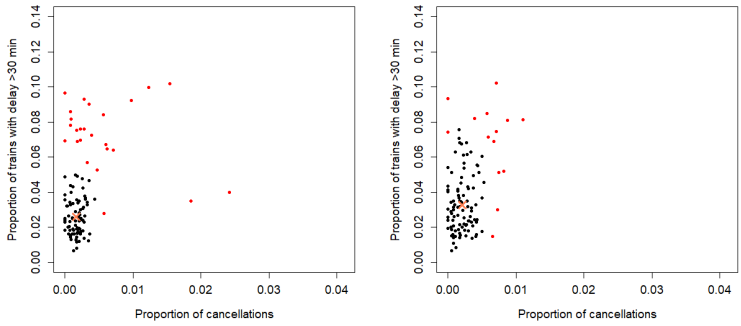
Given the newly obtained dataset, the analysis of Chapter 4 is repeated to obtain more reliable results.

6.1 Cancellations and extreme delays

Taking into account the percentage of canceled trains and the percentage of trains with a delay greater than 30 minutes, a multivariate robust analysis is repeated.
The robust estimates of the locations are recomputed and the obtained results are reported:

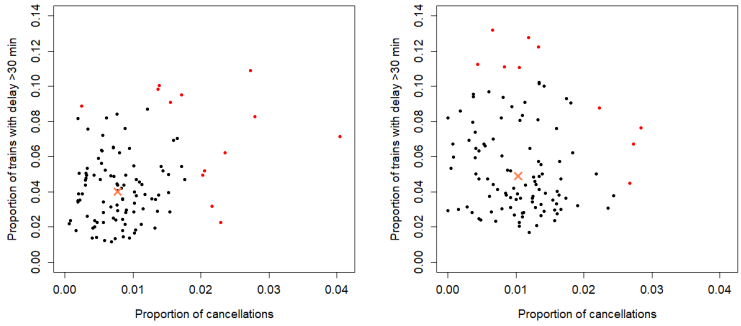
	% of canceled trains	% of delays >30 minutes
2015	0.2%	2.6%
2016	0.2%	3.3%
2017	0.8%	4.0%
2018	1.0%	4.9%

Compared with the results obtained in Table 4.1, the extreme delays trend does not seem to be impacted by the removal of the strike months. On the other hand, as expected, the canceled trains percentage dramatically decreased in 2016 and 2018. The results still show an increasing pattern over the years, although it is downsized.
Moving to the obtained outlying routes, the final result of the analysis is reported:



(a) Multivariate robust analysis in 2015

(b) Multivariate robust analysis in 2016



(c) Multivariate robust analysis in 2017

(d) Multivariate robust analysis in 2018

Figure 6.1: Outliers after the reweighting step are marked in red while the robust estimated location is marked with a cross

Route	2015	2016	2017	2018
Paris Lyon - Toulon	✓	✓	✓	✓
Nice Ville - Paris Lyon	✓	✓	✓	✓
Paris Lyon - Perpignan	✓	✓	✓	✓
Italie - Paris Lyon	✗	✗	✗	✓
Angouleme - Paris Montparnasse	✗	✗	✗	✓
Paris Lyon - Nice Ville	✓	✓	✓	✗
Lyon Part Dieu - Rennes	✓	✓	✓	✗
Paris Montparnasse - Toulouse	✓	✓	✗	✗
Toulouse - Paris Montparnasse	✓	✓	✗	✗
Toulon - Paris Lyon	✓	✓	✗	✗
Avignon - Paris Lyon	✓	✗	✗	✗
Rennes - Lyon Part Dieu	✓	✗	✗	✗
Nimes - Paris Lyon	✓	✗	✗	✗
Saint Malo - Paris Montparnasse	✓	✗	✗	✗
Lyon Part Dieu - Marseille	✓	✗	✗	✗
Paris Lyon - Annecy	✓	✗	✗	✗
Annecy - Paris Lyon	✓	✗	✗	✗
Valence Alixan - Paris Lyon	✓	✗	✗	✗
Paris Lyon - Montpellier	✓	✗	✗	✗
Montpellier - Paris Lyon	✓	✗	✗	✗
Lille - Marseille	✓	✗	✗	✗
Douai - Paris Nord	✗	✓	✗	✗
Paris Lyon - Saint Etienne	✗	✓	✗	✗
Paris Est - Strasbourg	✗	✗	✓	✗
Marseille - Lyon Part Dieu	✗	✗	✓	✗
Paris Est - Francfort	✗	✗	✓	✗
Stuttgart - Paris Est	✗	✗	✓	✗
Strasbourg - Paris Est	✗	✗	✓	✗
Marseille - Lille	✓	✗	✓	✗
Francfort - Paris Est	✓	✗	✓	✗
Perpignan - Paris Lyon	✓	✓	✗	✓
Lyon Part Dieu - Montpellier	✓	✓	✗	✓
Montpellier - Lyon Part Dieu	✓	✓	✗	✓
Lyon Part Dieu - Lille	✓	✗	✓	✓
Paris Lyon - Italie	✗	✓	✗	✓

Table 6.1: ✓ indicates the route is an outlier in that year, ✗ indicates it is not

From the bivariate plots, the identified outliers present at least one between high percentage of cancellations or extreme delays. Therefore, all the reported routes present some critical aspects over the period of observation. The Table 6.1 is colored according to the behavior of the outlying routes over the years.

- The most problematic routes are the first three, since they are detected over all the years
- The second most relevant are Italie - Paris Lyon and Angouleme - Paris Montparnasse, since they show a detriment in the performance, in particular in 2018
- The routes that show problems only at the beginning of the observation period are less critical because problems appear to be under control in the last time period
- The rest of observations cannot be categorized since no clear pattern can be identified

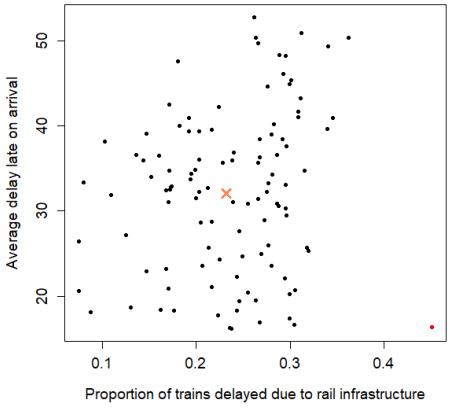
6.2 Causes

Since the goal of the project is to help the company improve its performance in the future, understanding the causes leading to a deterioration in the service can also be useful. For the sake of simplicity, in this section causes are treated as simple covariates, while in Chapter 7 an investigation over their compositional nature will be made.
A similar analysis to the one done in the previous paragraph is performed on the causes considering only the observations of 2018, given that those records are the most recent and can help the company in deciding how to further improve the service in 2019. Observations happened during the months with strikes have not been aggregated in the analysis.
The considered response variable in the analysis is the average delay at arrival of the trains arriving late, since it is the only variable leading to a consistent interpretation of the causes. Indeed, the causes in the dataset indicate proportions that are related to the trains that are late at arrival on a given route. Taking for instance the percentage of trains with a delay greater than 30 minutes as response would lead to inconsistent result, because the causes covariates are referred to all the trains arriving late and due to the aggregated nature of the dataset it is not possible to relate the causes with the trains with a delay greater than 30 minutes.
For each statistical unit, it is reported the proportion of times a delay has been caused by:

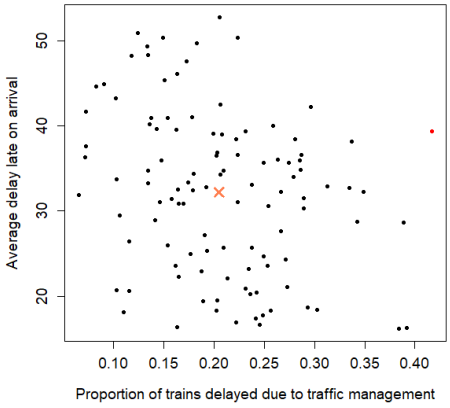
- Rail Infrastructure (RI)
- Traffic Management (TM)
- Rolling Stock (RS)
- Station Management (SM)
- Externals (E): it was decided to unify Travelers and the original Externals causes under the same part, since they are not directly imputable to SNCF.

A bivariate robust analysis is performed considering again the Minimum Covariance Determinant estimator with a reweighting step computed via FAST-MCD algorithm, setting $\alpha = 0.75$ and 2000 subsets for initial estimates. For each bivariate analysis a different cause is taken in consideration against the average delay at arrival.
The idea is to search for outlying routes that lead to high average delays related to a specific cause. Clearly the result would be only an indication, there is no causal inference that can be made since the results in the original dataset were already aggregated by month.
Results are summarized by the following table and the detailed plots are depicted.

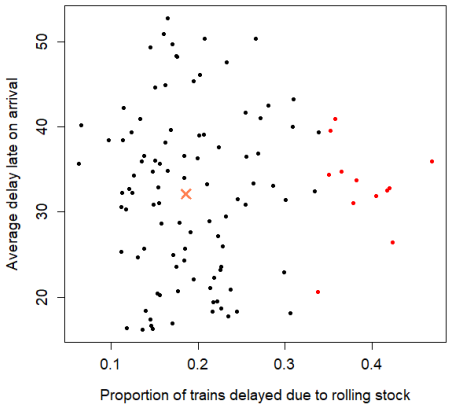
	Outliers	Outliers departing from Paris	Outliers arriving in Paris
RI	1	0	1
TM	1	0	1
RS	12	12	0
SM	3	2	1
E	0	0	0



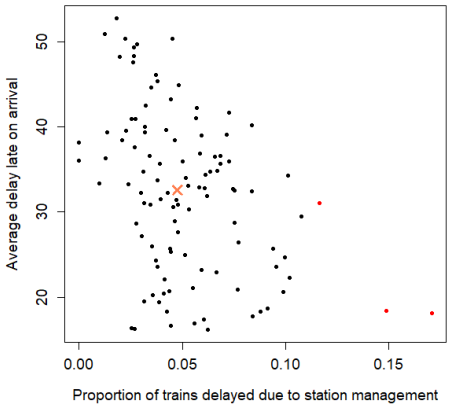
(a) Rail infrastructure effect



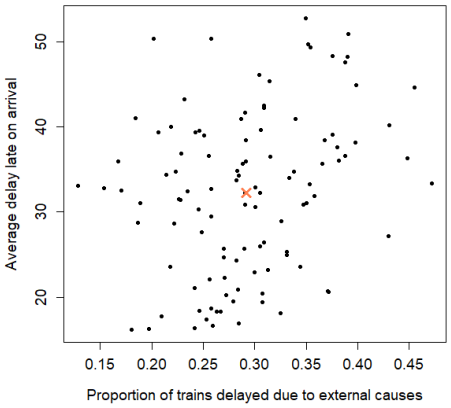
(b) Traffic management effect



(c) Rolling stock effect



(d) Station management effect



(e) External causes effect

Figure 6.2: Outliers after the reweighting step are marked in red while the robust estimated location is marked with a cross

As a result of the analysis, only weak conclusions can be made:

- The outlying routes that FAST-MCD algorithm detects are not related to their average delay at arrival. They show a high cause percentage in the bivariate plot, whatever it is the cause for which they are detected. This is not helping the company understanding which are the most problematic aspects of its performance, but it only highlights which routes are the most problematic due to a specific cause, even though their resulting delay might still be under control.
- All the routes identified as outliers depart from a Paris station or arrive in it. It is not a strong conclusion since the majority of the TGV routes involve a Paris station in the French railway system, but a further investigation will be made in Chapter 8.

The reported results are not really surprising, since the algorithm only identified problematic routes due to their cause of delay, no matter of the resulting delay time. A posteriori, this analysis could have similarly been performed with more simple methods, e.g. univariate boxplots.
Moreover, the causes are subject to a constraint that is not taken into account in the previous paragraph: they are non negative and must sum to 1, therefore they are constrained in a simplex and should be treated as compositional data.

6.3 Unifying the analysis

Given the nature of the multivariate outliers resulting from previous sections, finding routes that are detected from both the approaches would help the company understanding if a specific cause - Section 6.2 - may be behind the problems that result in a economic damage to SNCF - Section 6.1.
The main idea underlying the unification is that, even though no causal conclusion can be done, if a route is an outlier in 2018 for cancellation rates or extreme event rates and it also results as an outlier for a specific cause, the company has a better understanding of the type of needed intervention in order to improve its service. It must be highlighted that improving the performance due to a specific cause does not guarantee that the economic impacting events will decrease, given the aggregated nature of the original dataset.
Moreover, some differences can be highlighted since the economic impacting outlying routes are collected across the four years:

- If a route is an outlier along all the years, probably some systematic problems are present and having a better understanding of the main driving causes can have a positive impact on the company service.
- If a route starts becoming an outlier after some years of observations, problems may have arisen over the years and a prompt intervention on the driving causes could solve the problems before they become settled by time.
- If a route is not an outlier in 2018 or it has some oscillating behavior along the years, knowing the driving causes in 2018 does not help the company much.

Moving to the results, from the first part of the analysis in Section 6.1 five more critical outliers were detected. Among those, only the route Italie - Paris Lyon is also identified as an outlier by the bivariate robust analysis involving the causes, specifically it has an higher traffic management cause of delay with respect to the other routes. This route shows the highest percentage of delays over 30 minutes in 2018, thus an intervention on traffic management along the route may help solve the issue.
A weaker result is found for the route Paris Est - Strasbourg, which is detected from the economically impacting variables only in 2017, but not in 2018, and results as an outlier for the bivariate robust analysis involving rolling stock cause. However, this outcome does not suggest any possible improvement for the company, since this route is not critical in 2018.

Overall, the results of this chapter are giving a better understanding of which routes should be monitored closely in the following period by the company.

Compositional data analysis

As stated in Section 6.2, the causes percentages are subject to a constraint: they must be non negative and sum to 1, so they should be treated as compositional data objects in a 5-dimensional space. The causes percentages will be called parts in the current chapter.
For each statistical unit, it is reported the proportion of times a delay has been caused by:

- Rail Infrastructure (RI)
- Traffic Management (TM)
- Rolling Stock (RS)
- Station Management (SM)
- Externals (E): all the causes not imputable directly to the company behavior

Compositional data have some issues [11]:

- these kind of data do not belong to an Euclidean space ($\mathbb{R}^5$), yet to a 5-dimensional part composition (they are subject to the aforementioned constraints). As a result, the correlations are not free to range unrestrictedly anymore, being subject to a bias toward negative values;
- they present a marked curvature.

They can be represented in a graphical way:

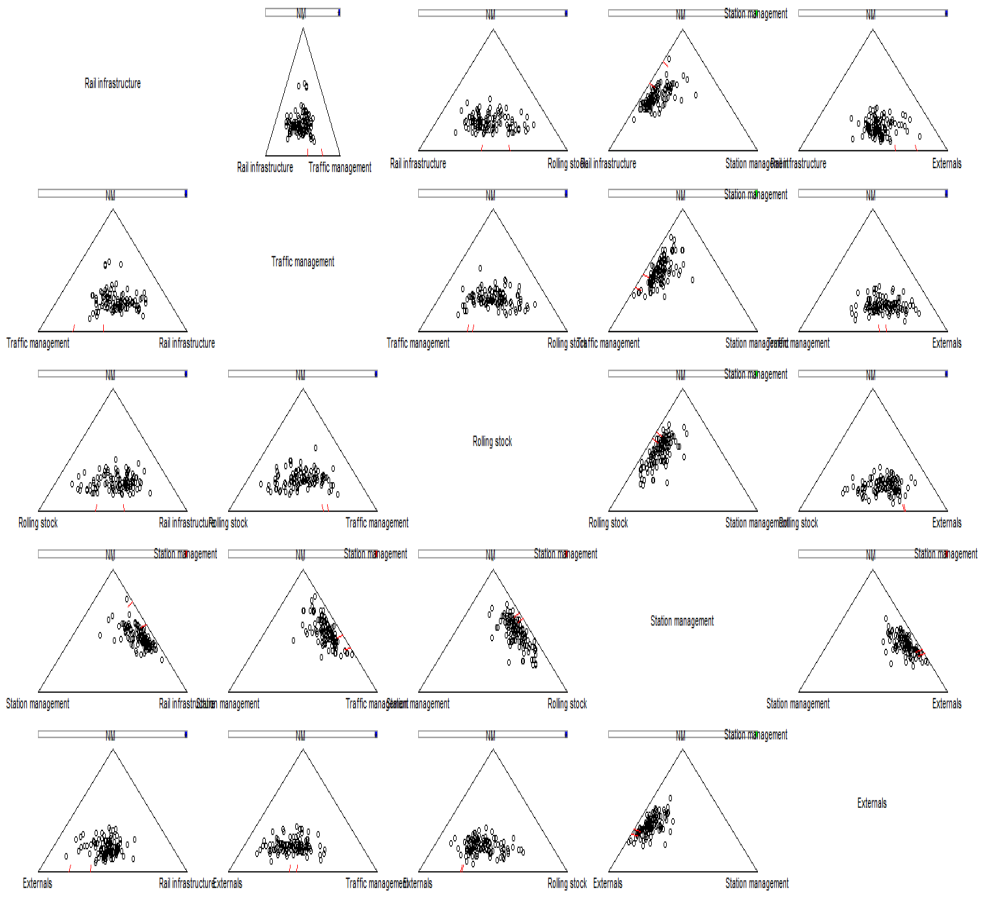


Figure 7.1: Parts of the causes in the Aitchison's simplex

7.1 Impact of the causes on the delays

The aim is to explain the response y, which is again the average delay at arrival of trains arriving late, as a function of all the causes. The goal is to try to analyze the effect of each part separately, in order to understand the total effect of each cause and spot the most impacting ones.

7.1.1 Methodology

In order to solve the issues arising from compositional data, as stated by Filzmoser et al. (2010) [12], it is possible to map compositional data from the simplex sample space to the usual Euclidean space, using one transformation belonging to the family of log-ratio transformations. One popular choice is the isometric log-ratio (ilr) transformation.
Starting from an observation $\mathbf{x} = (x_1, ..., x_d) \in S^d$, it is possible to recover

$$\mathbf{z} = (z_1, ..., z_{d-1}) \in \mathbb{R}^{d-1} \text{ s.t. } z_i = \sqrt{\frac{i}{i+1}} \ln\left(\frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}}\right), \qquad i = 1, ..., D-1.$$

Once the transformed variable is interpreted, a classical multivariate analysis - like regression - can be naturally performed using the new covariates. Taking the order of the parts as above, the following transformed covariates are obtained:

- $z_1 = \sqrt{\frac{1}{2}} \ln\left(\frac{RI}{TM}\right)$: since the logarithm is a monotonic function, it can be seen as how much the effect of RI is bigger than the one of TM;
- $z_2 = \sqrt{\frac{2}{3}} \ln\left(\frac{\sqrt{RI * TM}}{RS}\right)$: since the square root and the logarithm are monotonic functions, it can be interpreted as how much the effect of the interaction between RI and TM is bigger than the effect of RS;
- $z_3 = \sqrt{\frac{3}{4}} \ln\left(\frac{\sqrt[3]{RI * TM * RS}}{SM}\right)$: following the same reasoning of z_2, it can be interpreted as how much the interaction between RI, TM and RS is bigger than the effect of SM;
- $z_4 = \sqrt{\frac{4}{5}} \ln\left(\frac{\sqrt[4]{RI * TM * RS * SM}}{E}\right)$: for the same considerations on monotonicity, it represents how much the interaction between internal causes is bigger than the effect of the external causes.

The idea is to investigate the effect of the causes on the response by fitting a regression with the ilr transformed variables as covariates.
Thanks to the way the data has been pre-processed, statistical units can be considered independent.
Firstly, a linear model has been fitted:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \varepsilon$$

Although the gaussian assumption is satisfied at level 1% (the p-value of the Shapiro test on the residuals is 0.03267), the R^2 coefficient of the model is not high (0.197).
Fitting a Generalized Additive Model (GAM) appears to be a wise choice in order to be able to handle the 4 variables separately, using natural cubic B-spline basis to model possibly non-linear effects.

$$y = \beta_0 + f_1(z_1) + f_2(z_2) + f_3(z_3) + f_4(z_4) + \varepsilon \tag{7.1}$$

The R^2 of the model is only 0.219, while the residuals are also gaussian (the p-value of the Shapiro test is 0.06268). Actually, some terms appear to have a linear behavior: a semiparametric model can be more appropriate.

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + f_3(z_3) + f_4(z_4) + \varepsilon$$

The R^2 is 0.219, while the residuals remain gaussian. Inference on some coefficients was made to reduce the model considering the parametric test since the assumptions are met.

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

The resulting p-value is 0.062, and β_1 can be assumed to be null.
After removing the first covariate, it can be checked whether a second reduction is reasonable:

$$H_0 : \beta_2 = 0 \text{ vs } H_1 : \beta_2 \neq 0$$

The resulting p-value is 0.503, so β_2 can be considered null.
The reduced model is

$$y = \beta_0 + f_3(z_3) + f_4(z_4) + \varepsilon \tag{7.2}$$

The R^2 is 0.204. It is necessary to test if there is any difference in variance explained between the models 7.2 and 7.1: if not, the simplest model is retained.
H_0: the two models are equivalent vs H_1: the more complex model explains more variance
The test is done in a parametric fashion since the gaussian assumption is met for both models and the resulting p-value is 0.1751.
Furthermore, given the results on the reduced model, the term z_3 appears to have a linear behavior.

$$y = \beta_0 + \beta_3 z_3 + f_4(z_4) + \varepsilon \tag{7.3}$$

The test H_0: the two models are equivalent vs H_1: the more complex model explains more variance
is performed between the models 7.3 and 7.1 in a parametric setting, since the assumptions are met. The resulting p-value is 0.2194: there is no statistical evidence to reject H_0.
The R^2 of the chosen model 7.3 is 0.199, it has not decreased so much from the one of the original model, which was 0.219. The significance of the final semiparametric model is also justified by the test:

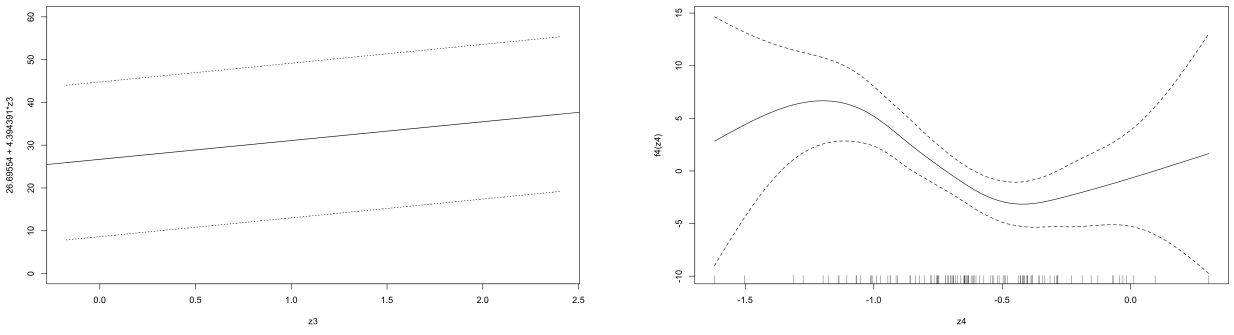
H_0 : model 7.3 and the null model explain the same amount of variance vs
H_1 : model 7.3 explains more variance

It has a p-value of 0.00006353, therefore 7.3 is a significant model.

7.1.2 Results

Results for the considered semiparametric additive model are reported

$$y = \beta_0 + \beta_3 z_3 + f_4(z_4) + \varepsilon$$



(a) Effect of z_3 (b) Effect of z_4

Figure 7.2: Semiparametric GAM for average delay late on arrival vs causes

The interpretation for the considered covariates is as follows:

- $f_3(z_3)$ is linear with positive slope and with residual standard error estimate 9.036:
  - If the interaction $RI * TM * RS$ is bigger than SM, the mean value of the response is bigger. Moreover, this term has a big impact in term of magnitude on the response;
  - The effect of term z_3 has a big effect on the magnitude of the response, also due to the high variance estimate. Furthermore, the introduction of RS in the interaction term is the one that made the biggest impact in the delay: Rolling Stock is the cause that has to be taken into account more seriously.
- $f_4(x_4)$ is a non-linear function:
  - It increases if $z_4 < -1$ or $z_4 > -0.5$: it means that the response increases if there is a large imbalance between internal causes - that the company can manage - and external causes - that is something that SNCF cannot handle;
  - The term z_4 can be seen as a way to balance the two types of causes: the argument of the logarithm has to be about 0.4 in order to reach the minimum of $f_4(x_4)$. Therefore, the better approach is to retain internal and external effects separately in order to make effective interventions. It is important to notice that the external causes are not predictable.

The Generalized Additive Model considered in this paragraph has some drawbacks:

- It has a small R^2: 0.199, which is just slightly greater than the one of the original linear model 7.1.1, but it is necessary to introduce non-linear terms in order to explain some effects.
- The error's estimate for the effect of z_3 is quite big.
- The interpretation of the transformed variables and of the model itself is not immediate and quite hard. The deduced results appear to be reasonable.
- The selected model cannot be used to make predictions, for example, on the first month of 2019. The model has been constructed in this way in order to analyze which causes drove more the average delay on the arrival along the last year.
- Ilr transformation depends strongly on how the columns are ordered. In order to check the goodness of the proposed model, different orderings of the columns were considered: a change in the order of the covariates results in different ilr transformations. With any permutation, the final model retained one term that includes RS and another that includes E, therefore they all lead to the same conclusions of the original model. The model written as in 7.3 was used for the interpretation for the sake of simplicity, but without loss of generality.

7.2 Conclusions

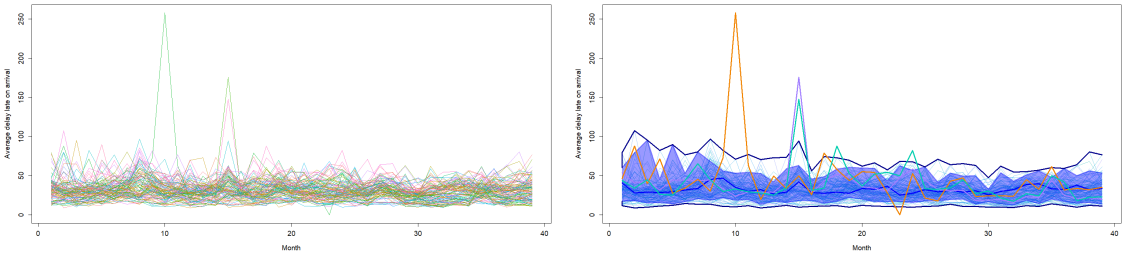
- Rolling stock seems to be the most critical issue to be managed by SNCF in order to improve its service in 2019.
- External causes are not manageable, but it is necessary for the company to be able to separate them from the internal ones: it should be able to control internal problems independently from the external conditions (weather, protests, strikes...) to make effective interventions. For example, it is important to manage well the traffic of the stations independently of the weather.

Final developments

After the analysis on the causes has highlighted some possible issues regarding Paris stations, a functional analysis about their influence is performed. From Section 6.2 they seem to be the most problematic in 2018, but a further investigation is necessary, also because Paris is the capital city of France and a lot of travelers visit it along the years.
First of all, it seems a better idea to consider the behavior of the time series along the 4 years, and taking a functional approach appears to be a reasonable choice. Instead of considering the whole time series, for the reasons discussed in Chapter 5 and to keep consistency within the analysis, only the months without strikes are considered.
A functional object is constructed, where each statistical unit is a route for which a variable is observed for 39 months.

8.1 Average delay late on arrival

The first variable considered is the average delay of the trains arriving late, which is also the response variable for Section 6.2.
By a preliminary analysis on the functional dataset, some amplitude outliers are detected by the functional boxplot: two of them arrive in Paris stations, while the third route is not involved with Paris. No shape outliers results from the outliergram.



(a) Representation of the functional data

(b) Functional boxplot

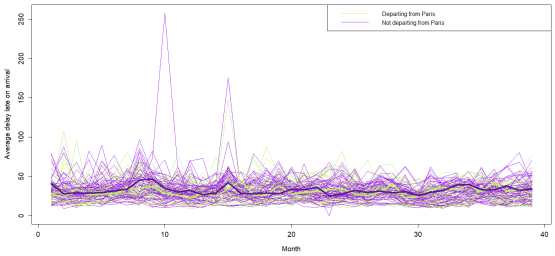
In the following, some statistical tests are performed in order to understand whether Paris stations have worse service than the other stations. In this chapter, the considered level of significance for tests is $\alpha = 0.05$.

8.1.1 Global test for Paris stations

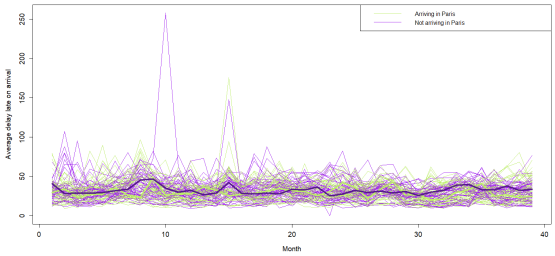
The functional approach was undertaken in order to perform a permutational nonparametric test for the influence of Paris stations on the average delay for late arriving trains.
Call X_{Paris} the average delay at arrival for trains departing from Paris and X_{Other} the average delay at arrival for trains departing from other cities. The following test is conducted:

$$H_0 : Median(X_{Paris}) = Median(X_{Other}) \qquad \text{vs} \qquad H_1 : Median(X_{Paris}) \neq Median(X_{Other})$$

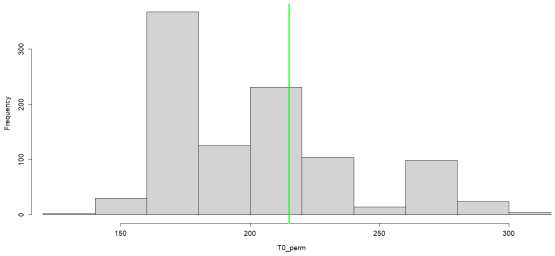
As test statistic, the L^∞ norm of the difference of the medians of the two groups is used.
By performing the test on the departure stations, the resulting p-value is 0.441. Therefore there is no statistical evidence to reject H_0.
By performing an analogous test on the arrival stations, the resulting p-value is 0.127. Therefore, again, there is no statistical evidence to reject H_0.



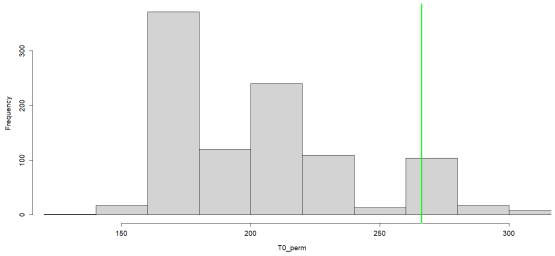
(a) Representation of the two groups, highlighting their medians - departure stations



(b) Representation of the two groups, highlighting their medians - arrival stations



(c) Permutational distribution - departure stations



(d) Permutational distribution - arrival stations

The conclusion is that the fact that a route starts or arrives in Paris has no statistical impact on the average delay at arrival.

8.1.2 Local test for Paris stations

Using the functional Benjamini-Hochberg procedure, the goal is to test whether the median average delay of trains arriving late is equal for trains departing from Paris and for those which are not. Each univariate test is performed using a permutational approach, using as test statistic the L^1 norm of the differences of the medians of the two groups.

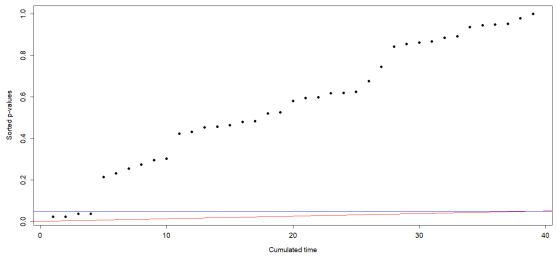


Figure 8.3: Analysis of the significance for Benjamini-Hochberg procedure

The result of the local test confirms the global one: in any observed month there is no statistical evidence to reject the null hypothesis: medians do not significantly differ in any month considering the departure stations.

The same test is performed on the arrival stations, comparing trains arriving in Paris and those with an arrival station in another city. The result is the following:

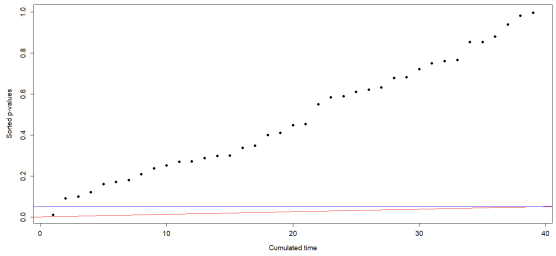
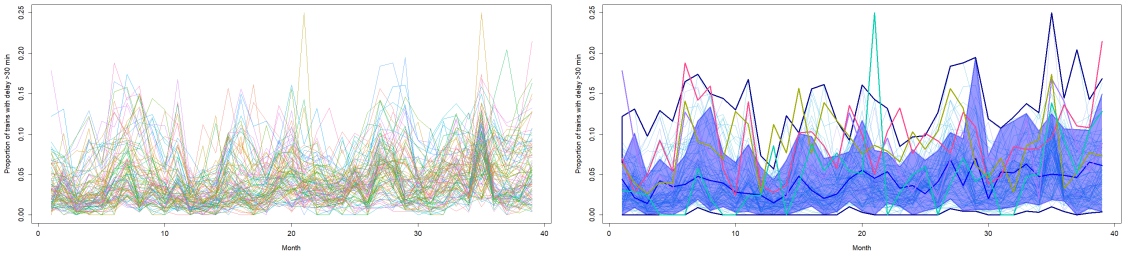


Figure 8.4: Analysis of the significance for Benjamini-Hochberg procedure

The result of the local test is coherent with the one of the global one: in any observed month there is no statistical evidence to reject the null hypothesis, meaning that medians do not significantly differ in any month also considering the arrival stations.

8.2 Proportion of trains delayed by more than 30 minutes

To understand if the analysis made in Section 6.1 are robust to a different approach, outliers detection can be performed in a functional fashion on the economic impacting covariates, i.e. canceled trains rate and extremely delayed trains.
Considering the extreme events rate as the variable, four amplitude outliers were discovered: routes Rennes - Lyon Part Dieu, Paris Lyon - Perpignan, Paris Montparnasse - Toulouse and Strasbourg - Nantes. Three of them were also detected in the corresponding bivariate analysis in Section 6.1, giving consistency to the analysis, although it is important to remark that now a univariate functional approach is employed. Moreover, two shape outliers are resulting from the outliergram, corresponding to the route Strasbourg - Nantes and its return trip.

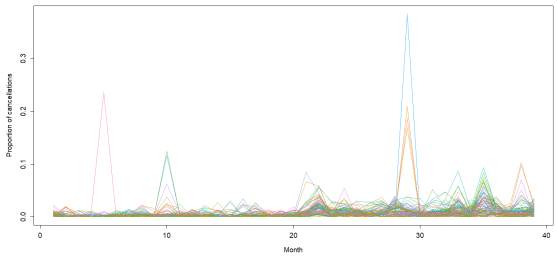


(a) Representation of the functional data

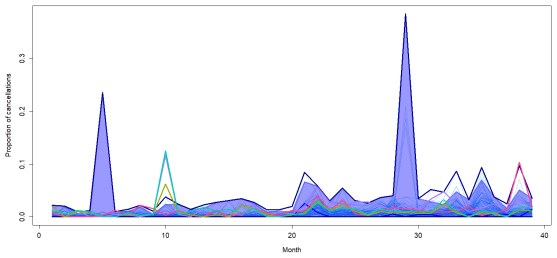
(b) Functional boxplot

8.3 Proportion of canceled trains

Considering instead the rate of canceled trains, no shape outliers are once again detected, while 4 amplitude outliers were discovered: Toulon - Paris Lyon, Montpellier - Lyon Part Dieu, Paris Lyon - Nice Ville and its return trip. All of them were also found in Section 6.1, adding consistency to the multivariate robust analysis since they are detected also from a different methodology.



(a) Representation of the functional data



(b) Functional boxplot

Conclusions

The performed analysis shows a detriment in the performance over the 4 years in terms of cancellations, delays and extreme events, i.e. delays over 30 minutes.
Some more detailed conclusions are presented. In terms of routes, It is possible to identify five routes on which a prompt intervention is suggested in order to limit the economic loss due to refunds:

- Paris Lyon - Toulon
- Nice Ville - Paris Lyon
- Paris Lyon - Perpignan
- Italie - Paris Lyon
- Angouleme - Paris Montparnasse

In particular, a possible driving cause is found for the route Italie - Paris Lyon. In this case, the suggested intervention is related to traffic management.
The cause having a greater impact on the average delay is found to be rolling stock. Therefore, the company should consider to invest in their rolling stock infrastructure, especially on the previously mentioned routes for which no main cause was detected.

To further enhance the analysis, a less aggregated dataset would be helpful. In particular, having data about each single trip of a train would enable to relate better the causes to the delays. Moreover, no information about intermediate stations of the routes was present in the dataset, but it could be useful to identify the most critical parts of the routes, especially for long routes.
A possible extension on the delay cause analysis would include the computation of Sobol' indices for the Generalized Additive Model to rank the transformed causes in terms of impact on the variability of the average delay at arrival. The additivity property of the model allows to interpret the effect of the inputs using only first order Sobol' indices.

Relevant code can be found in the GitHub repository: https://github.com/AndreaEnricoFranzoni/nonparametric_project

Bibliography

[1] Autorité de la qualité de service dans les transports (2018). *Les droits et démarches des voyageurs - Votre train est annulé.* URL: http://www.qualitetransports.gouv.fr/votre-train-est-annule-r172.html.

[2] SNCF (2018). *My train was delayed, can I claim compensation?* URL: https://www.sncf-connect.com/en-en/help/delay-train.

[3] The Guardian (3-04-2018). *French rail staff stage 'Black Tuesday' protests against overhaul.* URL: https://www.theguardian.com/world/2018/apr/02/france-mass-rail-strikes-macron-reforms-face-opposition.

[4] France 24 (8-03-2016). *Rail strike slows transport across France.* URL: https://www.france24.com/en/20160308-rail-strike-bring-travel-chaos-around-france.

[5] France 24 (25-04-2016). *French rail workers strike causes severe disruptions.* URL: https://www.france24.com/en/20160425-france-rail-trains-strike-severe-travel-disruption-sncf.

[6] Deutsche Welle (6-06-2016). *French president intervenes in train strike.* URL: https://www.dw.com/en/french-president-holds-talks-with-unions-to-avert-strike-action-during-euro-2016/a-19308452.

[7] The Washington Post (22-03-2018). *Strikes in France disrupt rail, air service in opening shot against Macron's labor plans.* URL: https://www.washingtonpost.com/world/europe/strikes-in-france-disrupt-rail-air-service-in-opening-shot-against-macrons-labor-plans/2018/03/22/ba0b1abe-8a69-4729-90c9-bd94d839f60c_story.html.

[8] La Tribune (28-08-2018). *Grève SNCF : le point sur les trains qui circulent ce lundi 28 mai (Transilien, RER, TGV et TER).* URL: https://www.latribune.fr/entreprises-finance/services/transport-logistique/greve-sncf-combien-de-trains-ce-lundi-28-mai-2018-779807.html.

[9] Independent (6-06-2018). *French rail strikes: British holidaymakers facing travel chaos as transport workers stage 48-hour walkout.* URL: https://www.independent.co.uk/travel/news-and-advice/france-rail-strikes-eurostar-tgv-cancelled-48-hour-walkout-transport-workers-latest-a8385436.html.

[10] Le Parisien (19-07-2018). *Grève SNCF : 4 TGV et TER sur 5 vendredi, premier jour de départ en vacances.* URL: https://www.leparisien.fr/economie/greve-sncf-4-tgv-et-ter-sur-5-vendredi-premier-jour-de-depart-en-vacances-05-07-2018-7808286.php.

[11] Aitchison, J. (1983). "Principal Component Analysis of Compositional Data". In: *Biometrika* 70.1, pp. 57–65.

[12] Filzmoser, P., Hron, K., Reimann, C. (2010). "The bivariate statistical analysis of environmental (compositional) data". In: *Science of The Total Environment* 408.19, pp. 4230–4238.