POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Project report

Waste Pollution in the United States of America

Author: Camilla Battistini, Michela Frigeri, Michael Ronzulli, Pietro Spina

Course: Nonparametric Statistics

Academic year: 2021 - 2022

---

1. Introduction

This project aims at helping no-profit associations that work hard everyday to clean our planet. These associations organize, periodically during the year,missions to clean up specific places (such as beaches, turistic monuments, streets etc.) but also expeditions to clean residual waste islands floating in the oceans. During these missions data are recorded, specifically the number of collected items (how many units of waste were collected during that mission), the number of total volunteers (how many people took part at the mission, helping collecting the garbage), the type of the mission (land cleaning or ocean cleaning, for example), a partial classification of recyclable items (Plastic/Foam or Glass/Rubber/Lumber/Metal) and other specific characteristics of each mission, such as the month, the day of the week, the association that organized the expedition etc.
Very often these associations are self-made organizations of few people that just want to make a difference improving their living environment by spontaneously taking part in this kind of expeditions. For this reason there might be some organization issues in the disposal of the collected waste, indeed usually during an expedition the volunteers collect all the waste in garbage bags and then all these bags are left to the organization behind the expedition to deal with. So at the end of each expedition the organization will have to manage the correct disposal and recycle of a huge amount of collected items, and in this framework it could be very useful to know a priori the expected amount of garbage we will likely collect during a future mission, according to the number of partecipants and the mission specific characteristics. Knowing the expected amount of waste *before* the start of the mission would be a useful tool for the organization, allowing the organizer to plan beforehand the practical disposal of garbage, taking appointment with some recycling center in order to deliver them a preannounced amount of waste in a specific day, or informing the local dumping sites that they will probably arrive with a consistent quantity of garbage, making sure that there is space for all those bags in order to avoid possible issues. To achieve this goal of organization improvement we want to build a prediction model at first for the totality of collected items, and secondly a model telling us how many of these items will be labeled as plastic, needing to be taken to a specific recycling center, obtaining as a consequence also the expected number of items made of "other" materials, that instead will be taken to the dump.
In the following sections is presented a first description of the dataset we will use during the whole project development, and then we will go further proceeding to analyze the data and looking for the better formulation of the above mentioned models.

Available Code

All the results presented in this report are obtained using the data and the code reported in the following GitHub repository: **NPS - Waste Pollution Project**.

The Dataset

The available websites to find data about environmental and pollution problems are many. In particular we picked up our data from **globalearthchallenge.earthday.org** since not only it is a reliable source, but it also provides a brief description of all the categorical variables and measures collected in the dataset, along with its *csv* and *shapefile* formats, and of course those informations were very useful for us to decide how to proceed through our analysis. The initial dataset was composed of 54.388 observations sourced from three citizen science marine litter projects, indeed this is an interoperable global plastics dataset from 2015 to 2018. Each observation consists of 78 attributes, either of *numeric*, *text* or *date/time* type, that we will not report here but that can be found by clicking on the website link already provided above. Of course to deal with such an amount of attributes would have been critical under various aspects, hence we selected just 17 of them which seemed to be the most relevant to model the phenomena. Our variables of interested are the following:

Attributes	**Description**
Location	Specific site in which the expedition took part
Country	Country in which the expedition took part
Subcountry	Subcountry of interest (for Federal States only)
Longitude	Spatial coordinate
Latitude	Spatial coordinate
Event Type	Type of expedition
Total volunteers	Number of volunteers that took part in the expedition
Year	Date in which the expedition took part
Month	Date in which the expedition took part
Day	Date in which the expedition took part
Day of the week	Day of the week corresponding to date
Total items	Number of collected items in the expedition
Plastic and foam	Percentage of plastic and foam items within the items collected
Glass, rubber, lumber and metal	Percentage of glass, rubber lumber and metal items
Continent	Continent of interest
Area	Extension of the subcountry in which the expedition took part

Note that in particular by *event type* we want to indicate one type of cleaning expedition between land cleanup, marine debris, underwater cleanup and watercraft cleanup. In this phase we also decided to remove some observations, due to the fact that we had unreadable or missing data, moreover despite the number of countries that we have available is large we see immediately that the proportion of data collected in the USA is clearly predominant with respect to the whole dataset, in particular they're 26549 over the total 40364 that remained after data cleaning. In order to get precise result from our analysis we focused only on the expeditions that took part in the USA, and thus we created a new separate dataset. Here you can find a shape file plot of the observations that we will try to model.

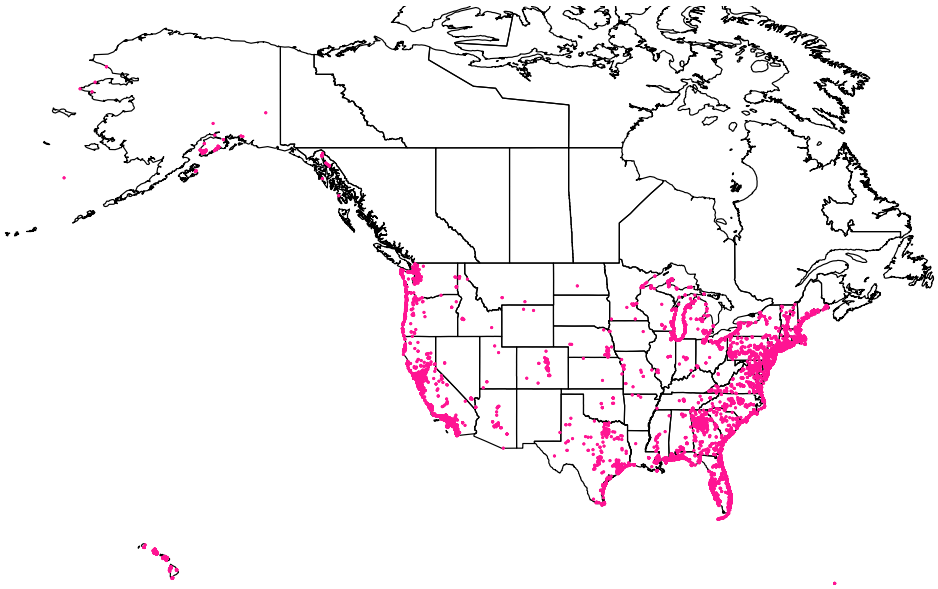

Figure 1: Observed Data in the USA

2. Explorative Data Analysis

2.1. Outliers Detection

First we computed some Depth analysis in order to detect multivariate outliers applying depth measures.
At first we explored the bivariate dataset composed by the only two numerical variables we have (Items vs Volunteers) using non parametric tools. Specifically we computed the Half-space (Tukey) Depth [HD] for our data points in order to identify possible outlying observations.

$$HD(F;x) = \inf_{H}\{P(H) : H \textit{ is a closed half-space in } \mathbb{R}^d \textit{and } x \in H\}$$

Below we provide the bagplot of such analisys from which we detected 9348 outliers.

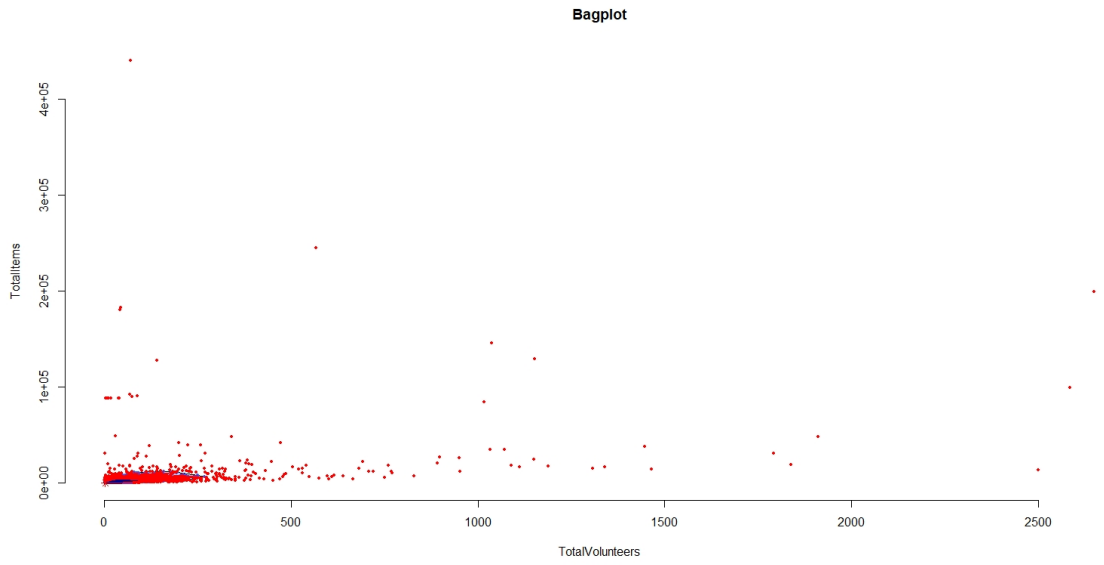

Figure 2: Bagplot Volunteers vs Items collected

Moreover we had a further confirm that those data are actually outliers by looking at the DDPlot, hence we finally decided to discard them from our analysis.

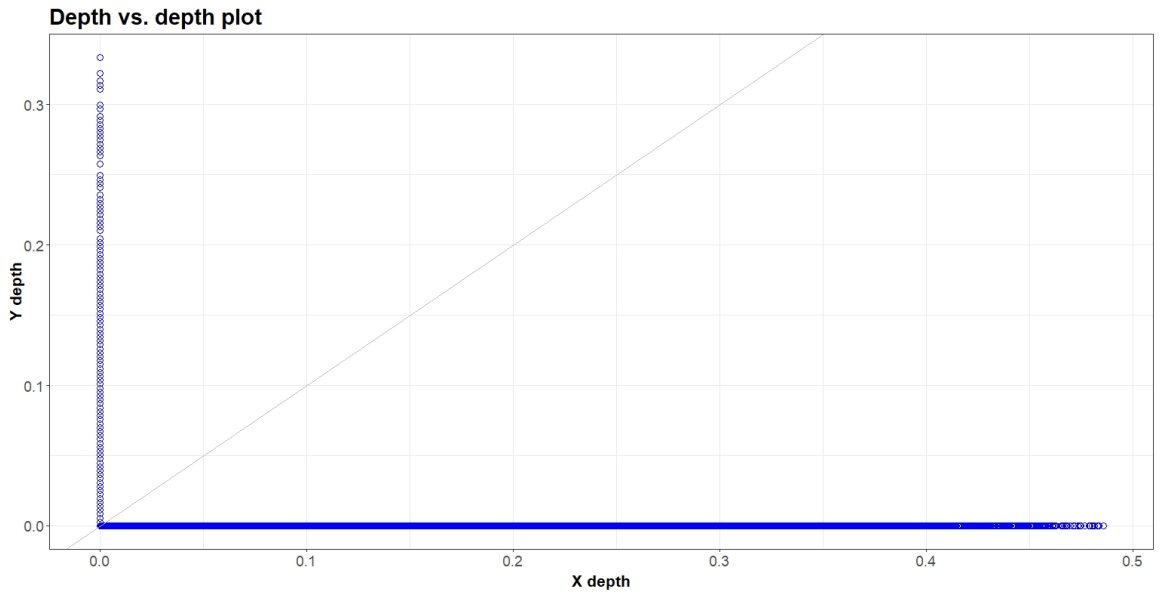

Figure 3: DDPlot of Outlying vs Not-Outlying observations

Depth measure techniques allowed us to significantly reduce the number of outliers present in our dataset but we went further by also making some choices based on simple common sense.
Indeed it is unrealistic that during a whole expedition one person doesn't pick up nearly any item or none at all, therefore we decided to exclude from our analysis also the observations for which the ratio between collected items and number of volunteers was less than five.

2.2. Data Transformation

One of our main goals is to make models and predictions about how many items will be collected during an expedition. We start by plotting on a bi-dimensional graph our observed data in order to try to guess which kind of relationship there is among them.
After a first look we decided to log-transform the number of items collected since in this way there seems to be a better shape for model fitting.
The cloud of data that we want to fit is hence the following:

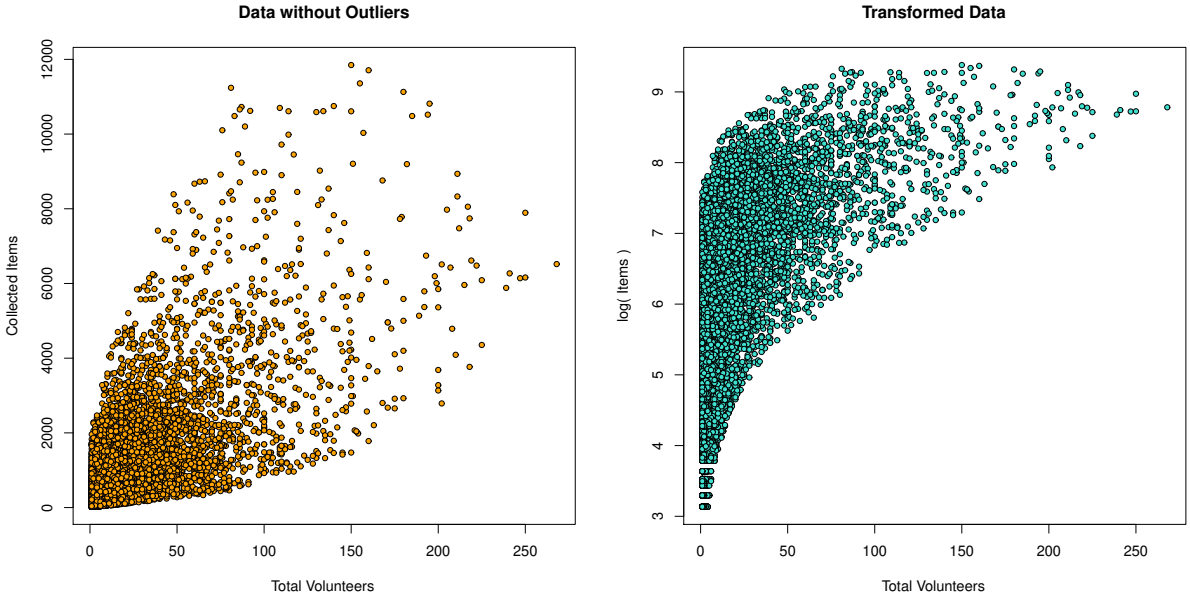

Figure 4: Data distribution after outliers removal, before and after the logarithmic transformation

So, from now on, with *TotalItems* we will refer to the log-transformed data.

3. Regression Model for Collected Items

3.1. Categorical Regressors

As the next step in our analysis we decided to try and implement different types of regression models in order to predict the logarithm of the total number of items collected in a location, given the number of volunteers taking part in the expedition and a collection of categorical variables.

From our dataset we already had the Event Type factor, that specifies the type of expedition. Its levels are:

- Land Cleanup
- Underwater Cleanup
- Watercraft Cleanup
- Marine Debris

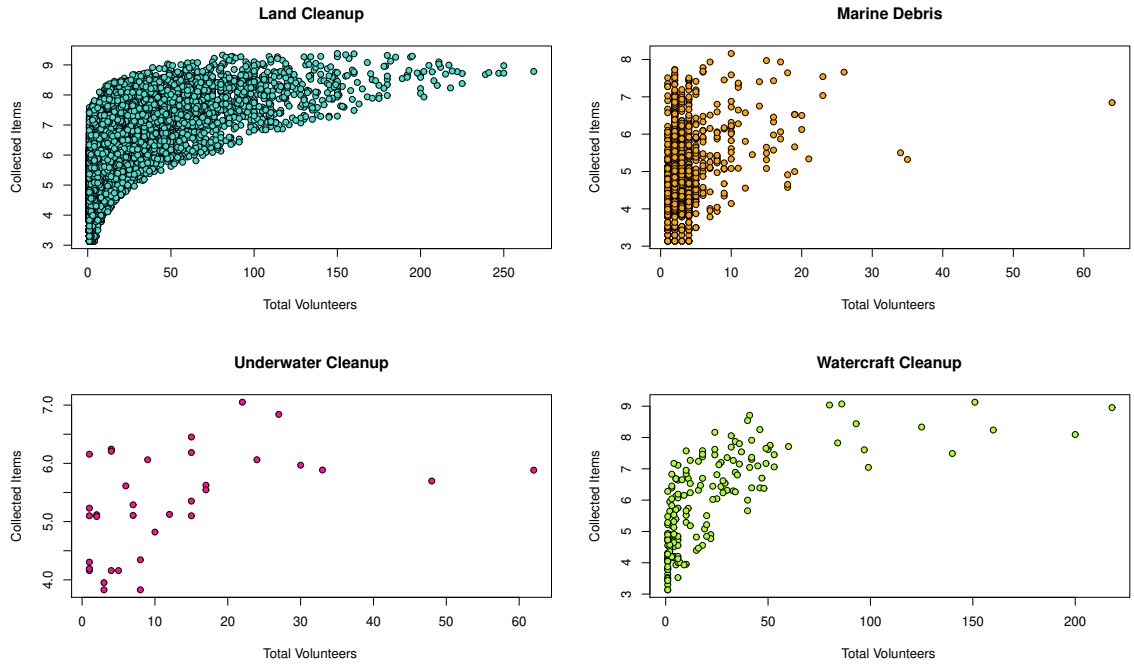

Figure 5: Regressor: EventType

We would also like to introduce a couple of new categorical covariates based on the time period of the expedition:

- *Season*: instead of considering each month as a categorical variable we decided to group them together in seasons: Winter, Spring, Summer and Autumn.

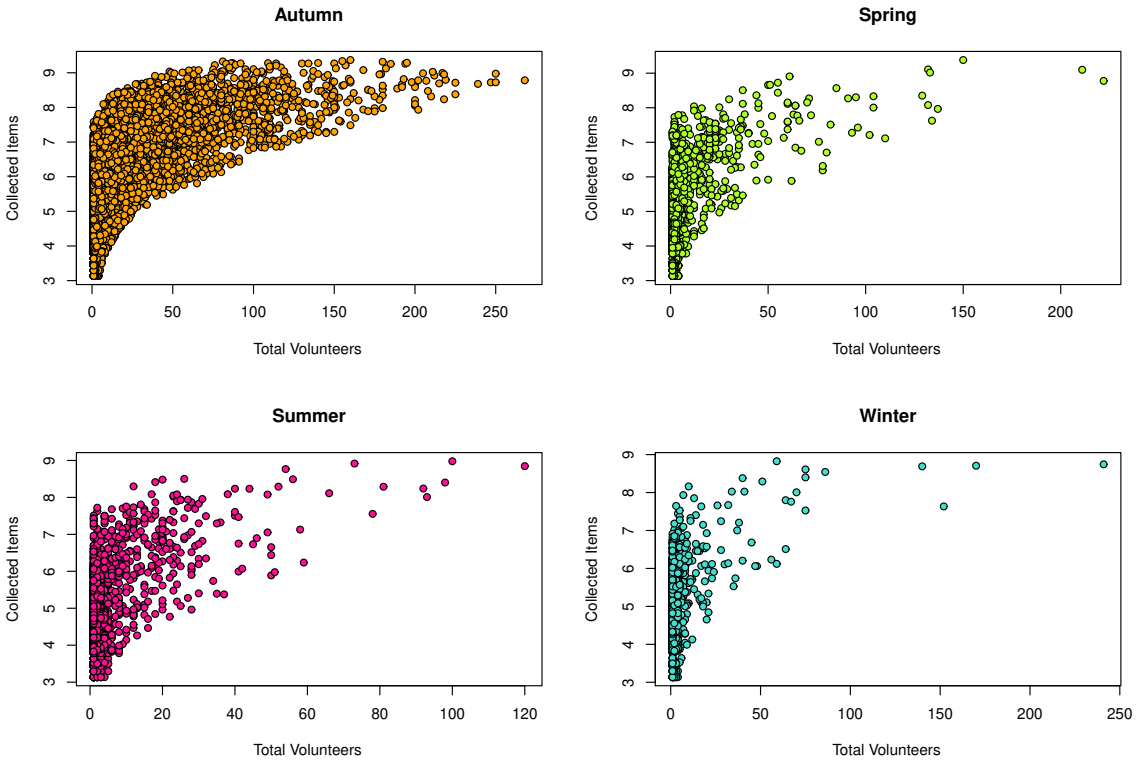

Figure 6: Regressor: Season

- *Weekday/Weekend*: instead of considering each day of the week as a categorical variable we decided to group them in weekday and weekend.

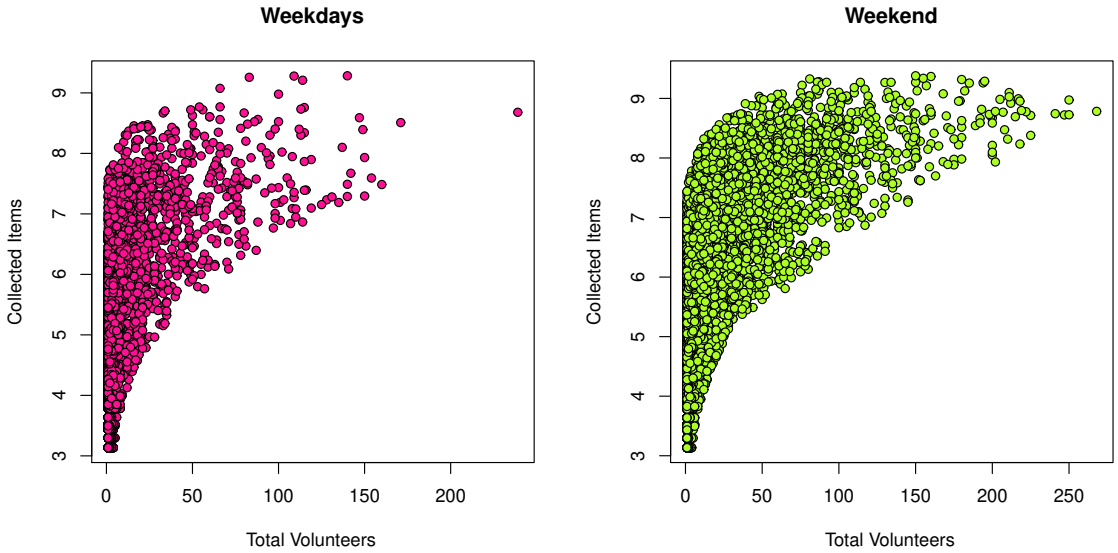

Figure 7: Regressor: Weekday/Weekend

3.2. Anova Test for Significance of Regressors

Since our data are not Gaussian the assumptions for ANOVA are violated, so we computed some Permutational Anova to understand which factors are significant and so which variables can be used as regressors for our model.

1. **One-Way Anova for EventType**

$$H_0 : X_1 \overset{d}{=} X_2 \overset{d}{=} X_3 \overset{d}{=} X_4 \text{ vs } H_1 : \exists i, j = 1, ..., 4 \text{ s.t. } X_i \overset{d}{\neq} X_j$$

p-value $= 0$ EventType **significant**

2. **One-Way Anova for Weekday/Weekend (Type of the day)**

$$H_0 : X_1 \overset{d}{=} X_2 \text{ vs } H_1 : X_1 \overset{d}{\neq} X_2$$

p-value $= 0$ Type of the Day **significant**

3. **One-Way Anova for Season**

$$H_0 : X_1 \overset{d}{=} X_2 \overset{d}{=} X_3 \overset{d}{=} X_4 \text{ vs } H_1 : \exists i, j = 1, ..., 4 \text{ s.t. } X_i \overset{d}{\neq} X_j$$

p-value $= 0$ Season **significant**

4. **Two-Ways Anova for EventType, Season and EventType:Season**

   - Test for EventType:Season (interaction between EventType and Season)

$$H_0 : X_1 \overset{d}{=} X_2 \overset{d}{=} ... \overset{d}{=} X_{16} \text{ vs } H_1 : \exists i, j = 1, ..., 16 \text{ s.t. } X_i \overset{d}{\neq} X_j$$

     p-value $= 0$ Interaction **significant**

   - Test for EventType

$$H_0 : X_1 \overset{d}{=} X_2 \overset{d}{=} X_3 \overset{d}{=} X_4 \text{ vs } H_1 : \exists i, j = 1, ..., 4 \text{ s.t. } X_i \overset{d}{\neq} X_j$$

     p-value $= 0$ EventType **significant**

   - Test for Season

$$H_0 : X_1 \overset{d}{=} X_2 \overset{d}{=} X_3 \overset{d}{=} X_4 \text{ vs } H_1 : \exists i, j = 1, ..., 4 \text{ s.t. } X_i \overset{d}{\neq} X_j$$

     p-value $= 0$ Season **significant**

In conclusion for our model we will take in account EventType, Season, EventType:Season and Weekday/Weekend as regressors.
If we choose $\alpha = 0.05$ we have:
TotalItems $\sim$ **EventType** $+$ **Weekday/Weekend** $+$ **Seasonality** $+$ **EventType:Seasonality**

3.3. Model Choice

In the following section we will introduce and discuss the performances of four different regression models. At the beginning we decided to consider only the quantitative variable *Total volunteers* to identify the best regression model, before including the different categorical variables in order to allow shifting.
We will now present four different models, their theoretical properties and discuss their goodness of fit.

3.3.1 Step functions regression

At the beginning we considered a Step function model with ten even bins:

$$y_i = \beta_0 + \beta_1 c_1(x_i) + \beta_2 c_2(x_i) + ... + \beta_{10} c_{10}(x_i) + \epsilon_i$$

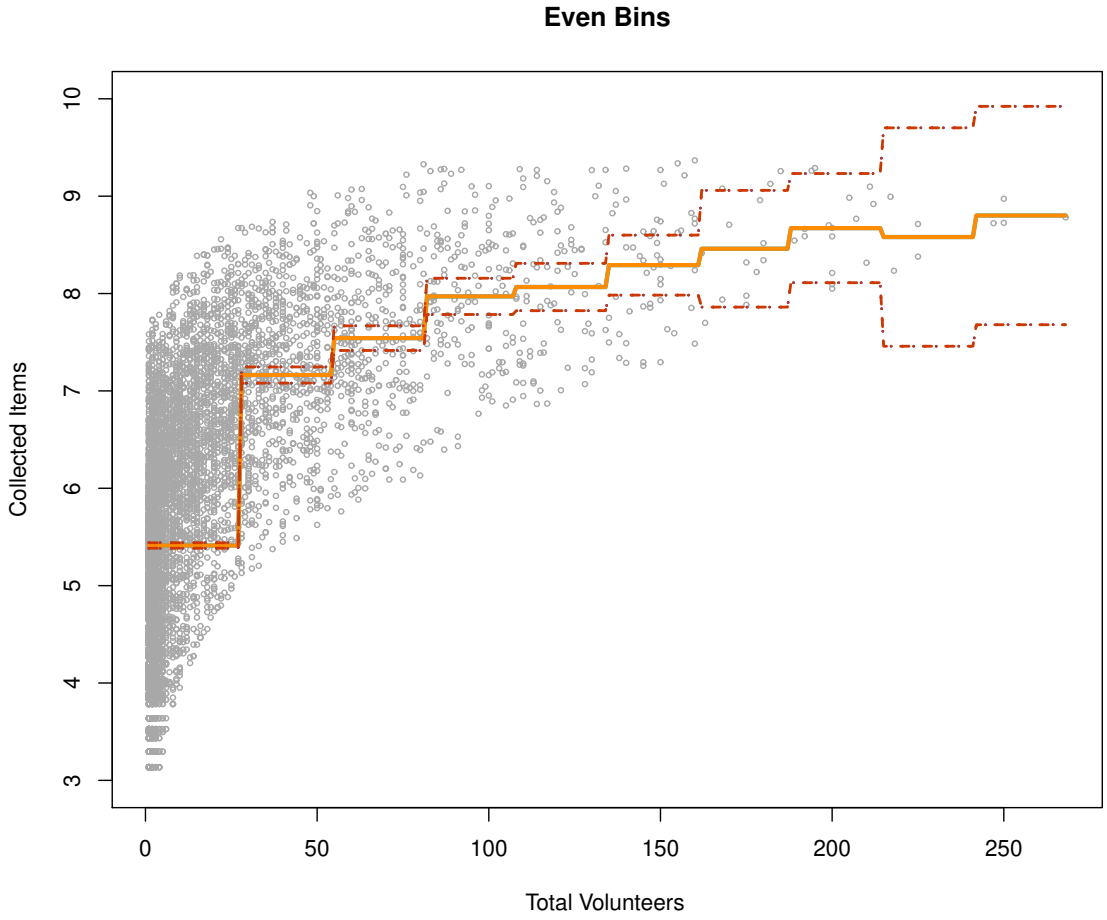

Figure 8: Step functions regression model with 10 even bins

The model doesn't seem to work to well indeed it has:

$$R^2_{adj} = 0.343$$

Which is fairly low, this is mainly because data are concentrated towards small values of *Total Volunteers*, indeed we have that the first bin:

$$c_1 = \mathcal{I}(0 \leq x_i \leq 27.7)$$

contains 6246 data points which is over 80% of the whole dataset.

To fix this issue we decided to introduce a Step functions regression model considering uneven bins with an higher concentration towards small values of *Total voluteers* as follows:

Bins	Number of data points
(0,2]	2978
(2,5]	1225
(5,10]	759
(10,20]	886
(20,30]	530
(30,40]	313
(40,50]	227
(50,75]	326
(75,100]	159
(100,300]	206

As we can see in the plot below the previous issue has less influence on the model compared to the even bins case:

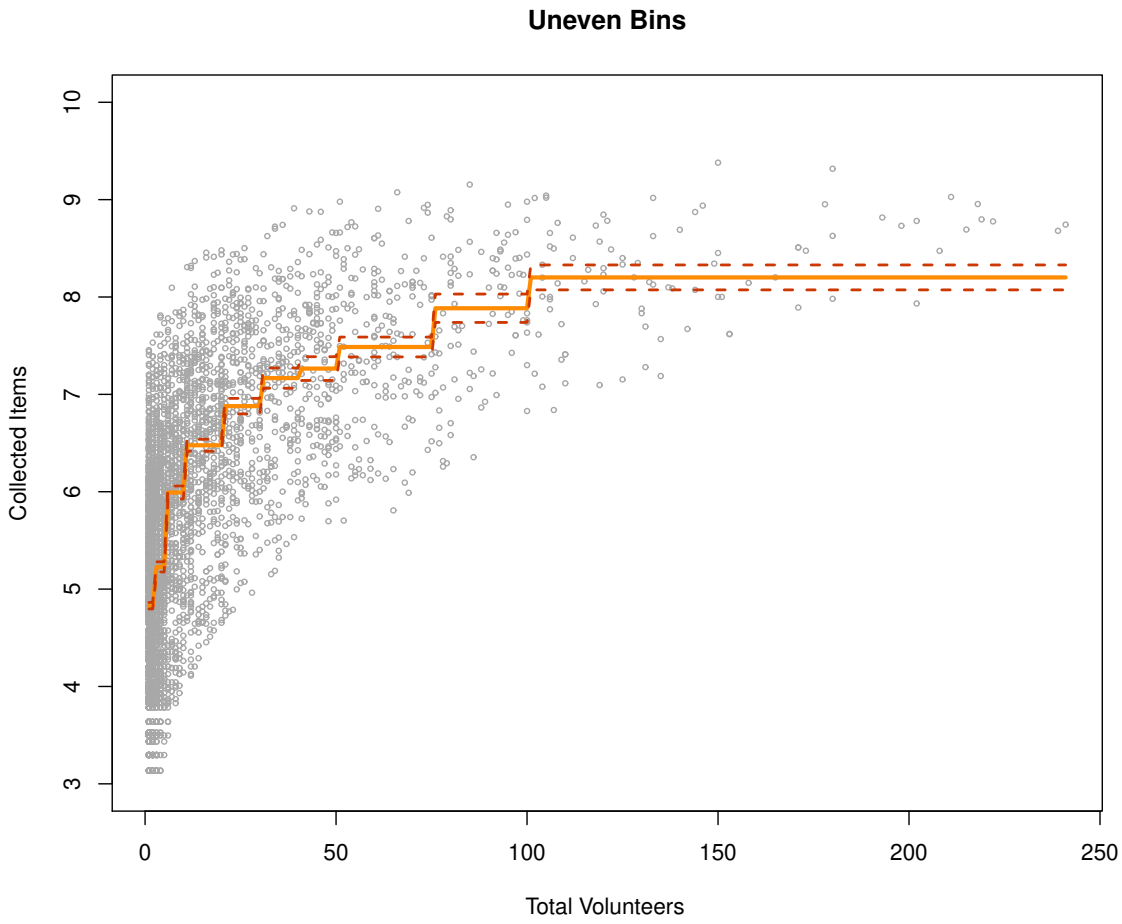

Figure 9: Step functions regression model with 10 uneven bins

As we can see the model improves significantly after this procedure

$$R^2_{adj} = 0.559$$

This model is also decently performing in terms of goodness of fit, as we can see from the below plots there are no significant leverage points, the plot of *Residuals vs Fitted* also performs pretty well, the only one non performing well is the *qqplot*, in fact if we perform an *Anderson-Darling test* to test the normality of the model's residuals we obtain $p\text{-}value = 0$, confirming the non-normality of the residuals.

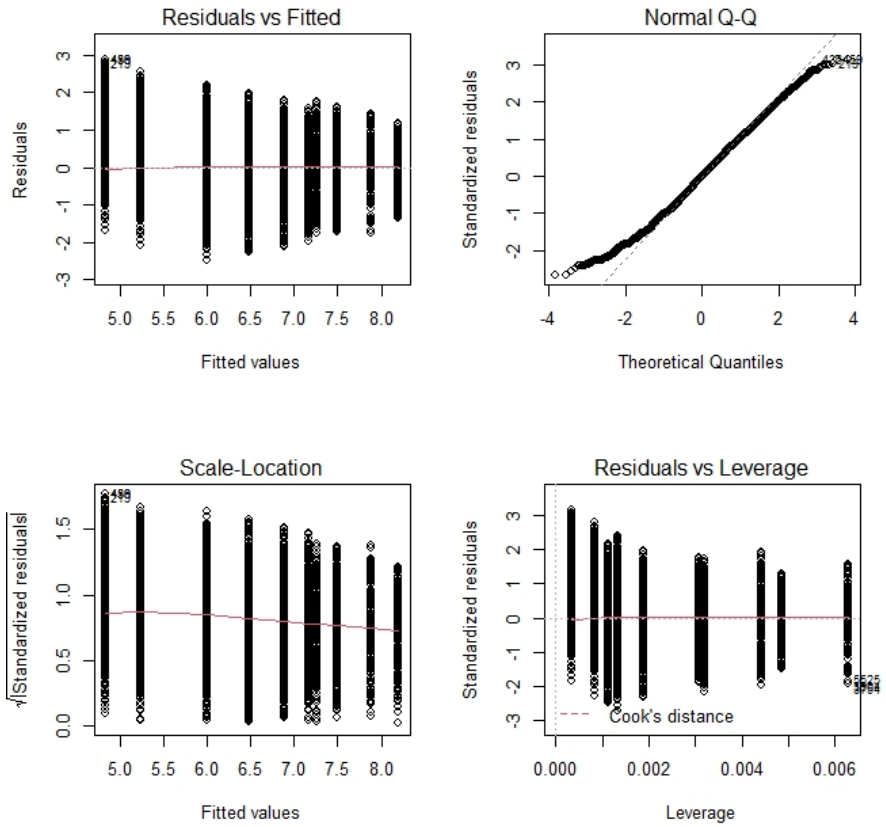

Figure 10: Step functions regression model with 10 uneven bins: Residuals

3.3.2 Polynomial regression

As the next step in our analysis we decided to consider a polynomial regression model.
We started by performing an Anova F-test comparing polynomial regression models from degree 1 up to degree 10, in order to find the optimal degree for the polynomial regression:

```
Analysis of Variance Table

Model  1: y ~ poly(x, degree = degree)
Model  2: y ~ poly(x, degree = degree)
Model  3: y ~ poly(x, degree = degree)
Model  4: y ~ poly(x, degree = degree)
Model  5: y ~ poly(x, degree = degree)
Model  6: y ~ poly(x, degree = degree)
Model  7: y ~ poly(x, degree = degree)
Model  8: y ~ poly(x, degree = degree)
Model  9: y ~ poly(x, degree = degree)
Model 10: y ~ poly(x, degree = degree)
   Res.Df    RSS Df Sum of Sq         F    Pr(>F)
1    7607 8915.3
2    7606 7492.9  1   1422.42 1721.3355 < 2.2e-16 ***
3    7605 6875.8  1    617.07  746.7492 < 2.2e-16 ***
4    7604 6563.9  1    311.92  377.4681 < 2.2e-16 ***
5    7603 6408.1  1    155.82  188.5713 < 2.2e-16 ***
6    7602 6330.9  1     77.21   93.4410 < 2.2e-16 ***
7    7601 6295.9  1     34.98   42.3317 8.192e-11 ***
8    7600 6283.5  1     12.42   15.0295 0.0001067 ***
9    7599 6281.1  1      2.42    2.9311 0.0869273 .
10   7598 6278.6  1      2.49    3.0149 0.0825434 .
```

Figure 11: Anova F-test p-values

As we can see from the p-values polynomials are influential to the model up to degree 8, therefore we proceed by creating a polynomial regression model of the 8th degree:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_8 x_i^8 + \epsilon_i$$

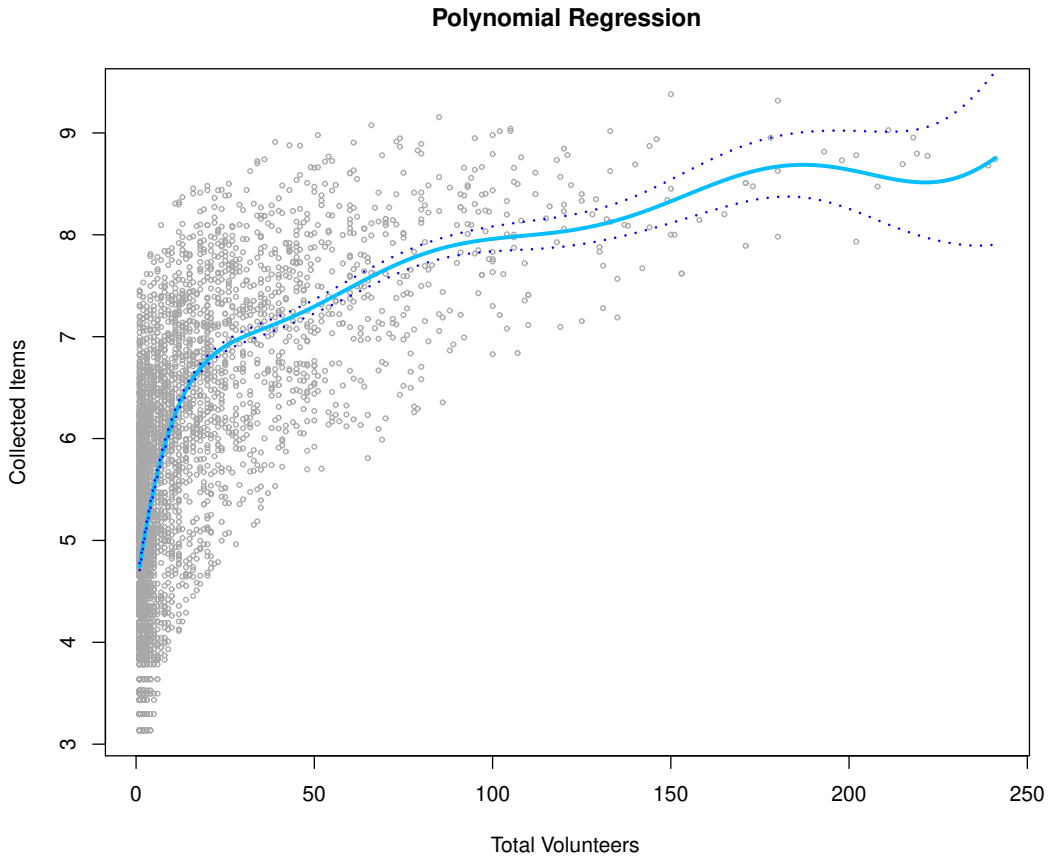

Figure 12: Polynomial regression of degree 8

As we can see the model seems to performs decently in terms of adjusted R-squared:

$$R^2_{adj} = 0.568$$

But if we look at the goodness of fit plots everything seems to go smoothly except for a very high leverage point (datum 2989), this is because this particular observation presents an extremely high number of *Total Volunteers* and relatively low number of *Total Items* picked up given how many volunteers were sent in that specific location. Therefore we proceed by removing this observation and we obtain a pretty good polynomial fit (Figure 12).

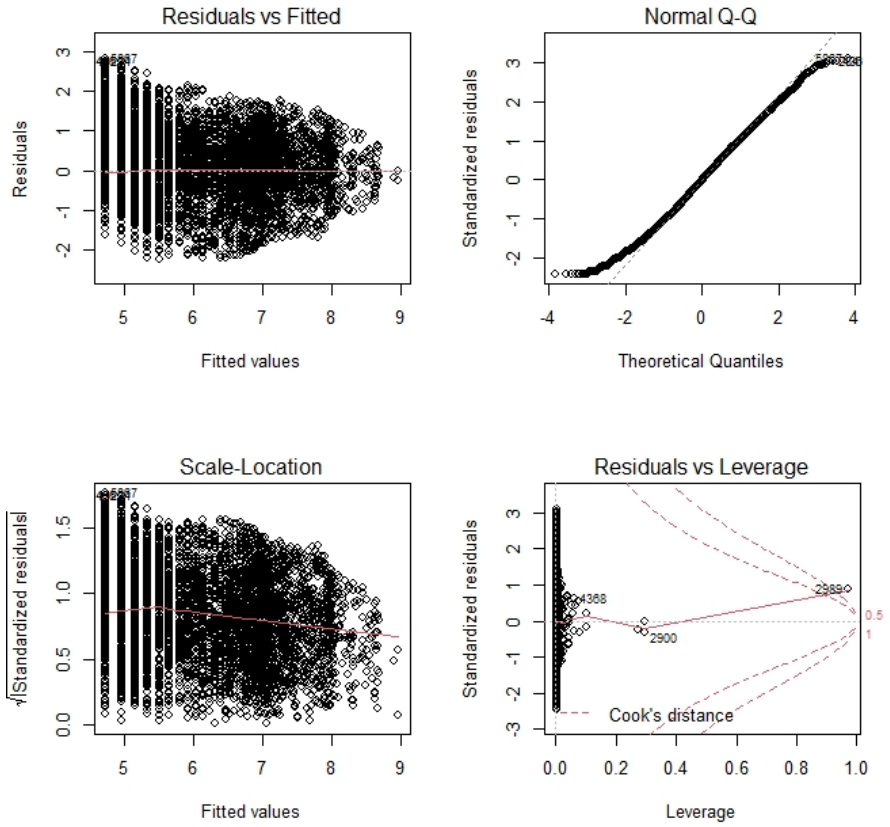

Figure 13: Polynomial regression of degree 8: residuals

3.3.3 Gaussian kernel regression

As the next step in our analysis we decided to consider a Gaussian kernel regression model (we also tried considering a uniform kernel but we obtained a really bad performing model due to lack of data in the bins).

$$\hat{y_i}|x_0 = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where :

$$w_i = \frac{1}{\sqrt{2\pi}} e^{\frac{(x_i - x_0)^2}{2h^2}}$$

In this case we consider a Gaussian kernel regression model with bandwidth set to 5 since after different trials it seemed to produce a smooth curve while maintaining an increasing trend.

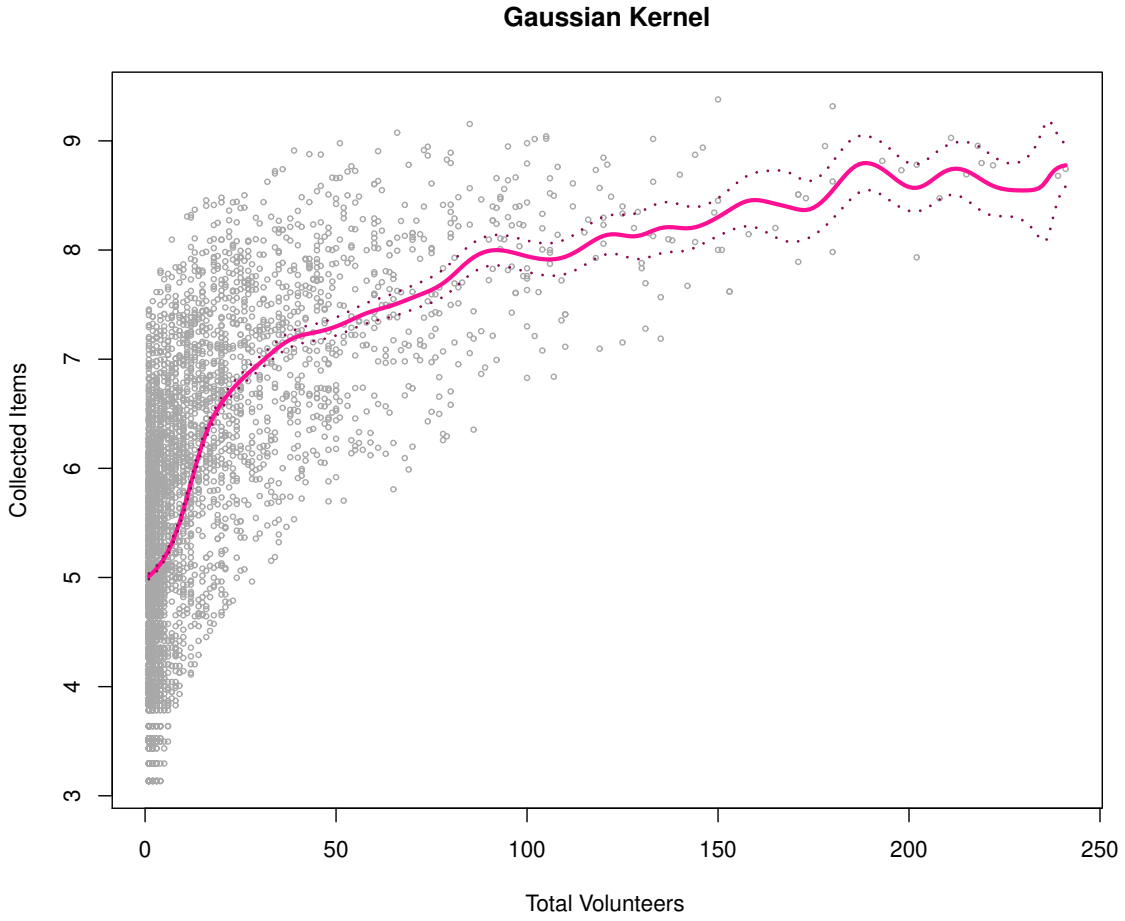

Figure 14: Gaussian Kernel

The model has a pretty good performance with decent adjusted R-squared:

$$R^2_{adj} = 0.532$$

3.3.4 Natural splines regression model

As the final step in our regression model analysis we considered a natural spline regression model setting the knots through a quantile driven approach, setting boundary knots at the 1% and a the 99% quantiles.

$$y = \beta_0 + \sum_{j=1}^{3+m} \beta_j g_j(x) + \epsilon$$

where,

$$g_j(x) = x^j \quad j = 1, 2, 3$$
$$g_j(x) = (x - t_i)_+^3 \; j = 4, .., 3 + m \; , i = 1, .., m$$

where m = 12 is the number of knots considered.

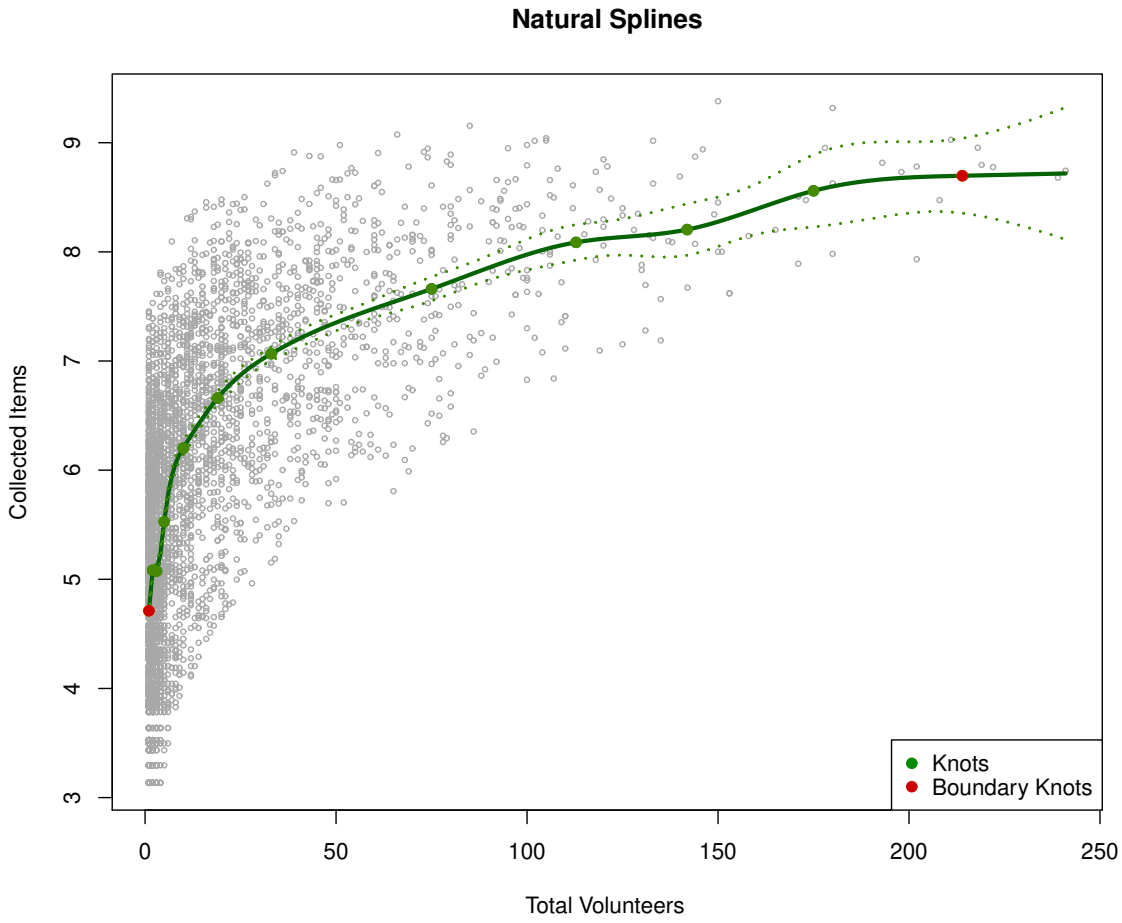

Figure 15: Natural Splines

The model has one the best performances so far, not only having a pretty high adjusted R-squared:

$$R^2_{adj} = 0.571$$

But also a pretty good behaviour in term of goodness of fit as we can see from the plots below, the only observation worth noting is the presence of a leverage point but since it doesn't strongly affect the model we can consider it as "good leverage point" and there is no reason to remove it from the analysis.

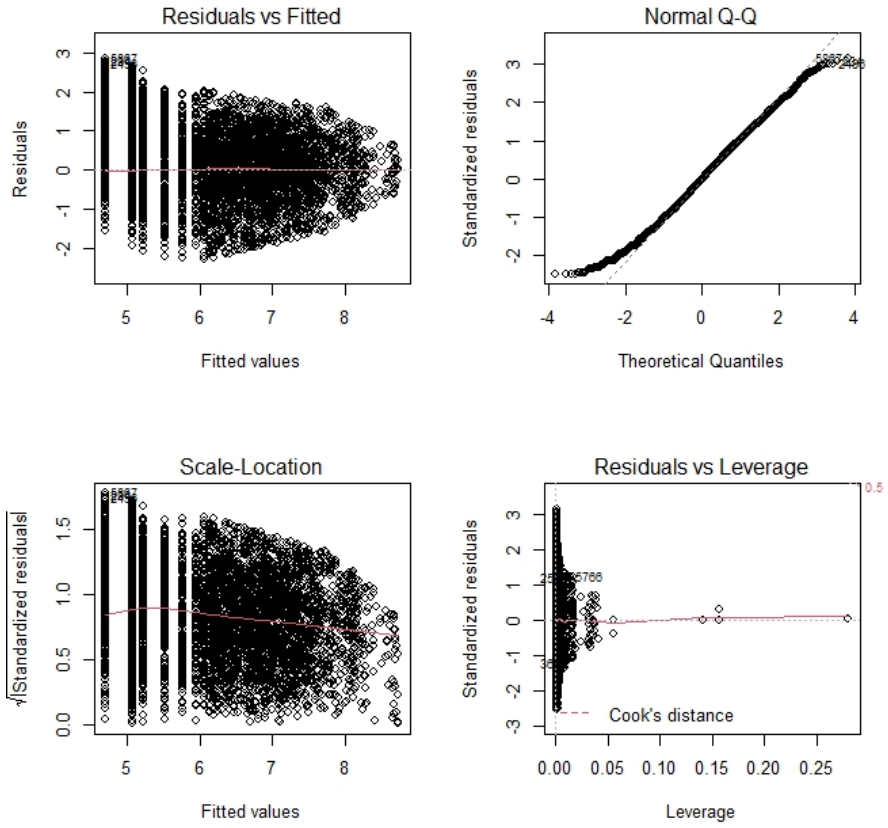

Figure 16: Residuals of the model

3.4. Final Model

At this point in our analysis we are ready to build the complete model with all our regressors (i.e. including the categorical regressors introduced in section 3.1).
Moving forward we decided to focus on the Natural Splines and Polynomial regression models given their superior performances when compared with the other two models presented in the section above.
In order to later perform conformal prediction on our models (section 4), and also to provide a visual effect of the shift caused by the intoduction of the categorical variables in our models we decided to split our dataset as follows:

- Training set: All data collected in the years 2016 and 2017
- Test set: All data collected in the year 2018

3.4.1 Natural splines regression model

$$y = \beta_0 + \sum_{j=1}^{3+m} \beta_j g_j(x) + \underline{\beta_{16}} * Weekday/Weekend +$$
$$\underline{\beta_{17}} * EventType + \underline{\beta_{18}} * Season + \underline{\beta_{19}} * EventType : Season + \epsilon$$

where :

$$g_j(x) = x^j \quad j = 1, 2, 3$$
$$g_j(x) = (x - t_i)_+^3 \; j = 4, .., 3 + m \; , i = 1, .., m$$

where m = 12 is the number of knots considered.

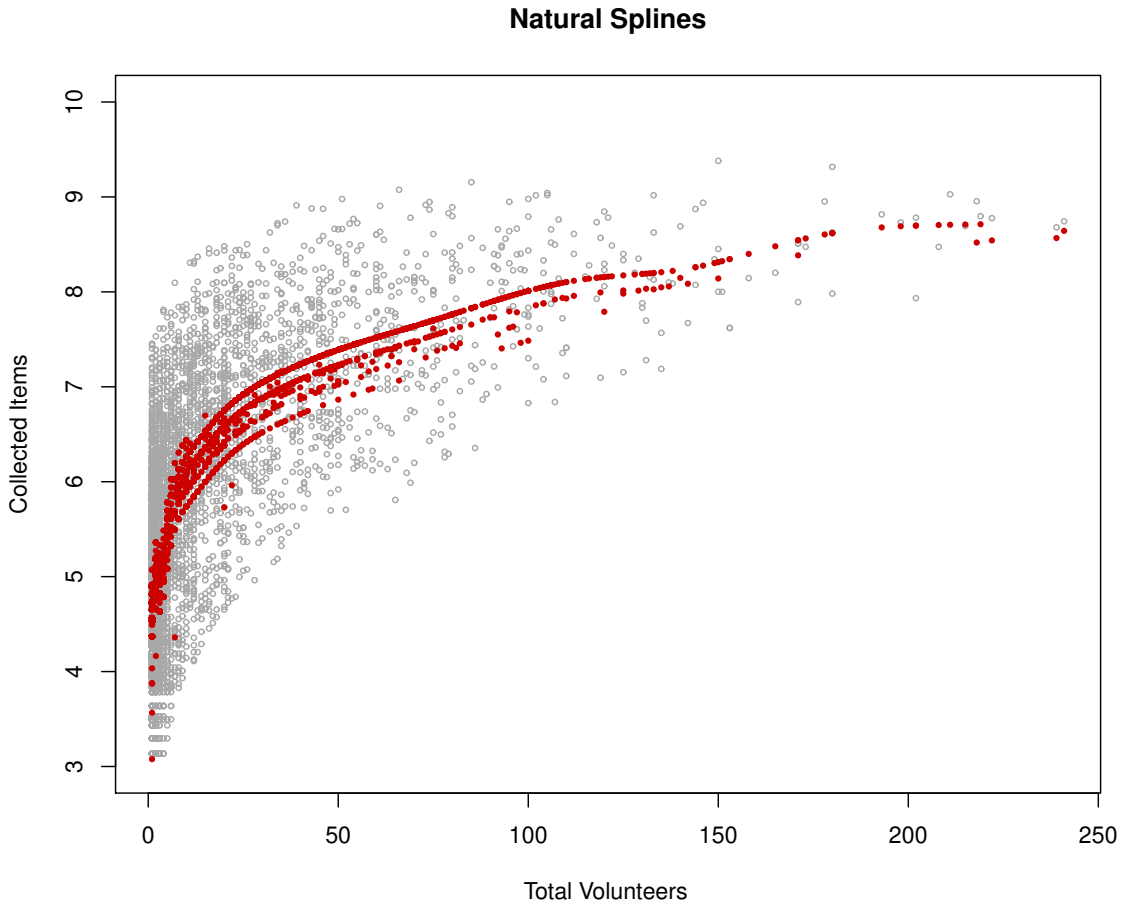

Figure 17: Natural Splines model

The R-squared and the adjusted R-squared are:

- $R^2 = 0.582$
- $R^2_{adj} = 0.580$

Moreover, having the chance to compare our predicted values with the true collected ones, we computed the *Root Mean Square Error*(RMSE) with the following formulation:

$$RMSE = \sqrt{MSE(\hat{y}_i)} = \sqrt{\mathbb{E}[(\hat{y}_i - y_i)^2]} = \sqrt{\frac{\sum_{k=1}^{N}(\hat{y}_i - y_i)^2}{N}}$$

obtaining $RMSE_{NS} = 0.91$

3.4.2 Polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_8 x^8 + \underline{\beta_9} * Weekday/Weekend +$$
$$\underline{\beta_{10}} * EventType + \underline{\beta_{11}} * Season + \underline{\beta_{12}} * EventType : Season + \epsilon$$

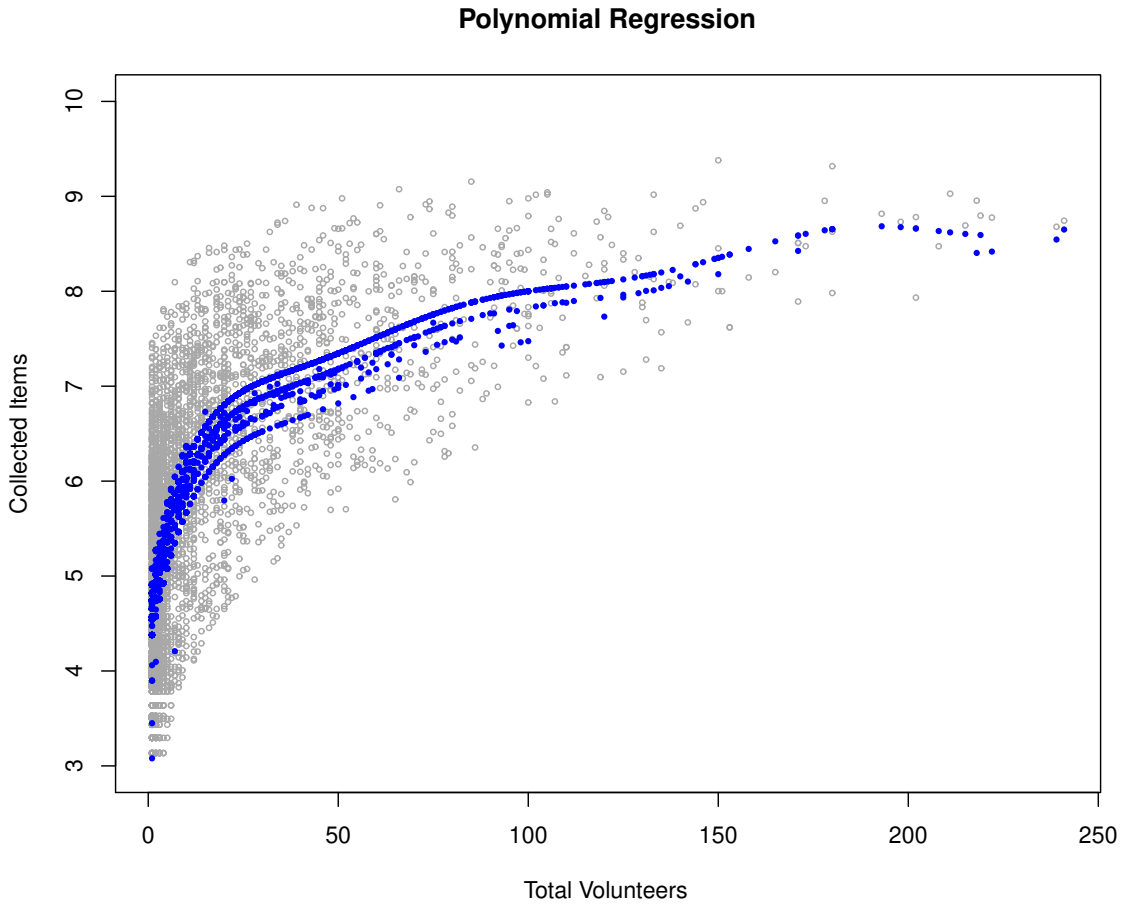

Figure 18: Polynomial regression model of degree 8

The R-squared and the adjusted R-squared are:

- $R^2 = 0.5795$
- $R^2_{adj} = 0.5781$

Also in this case we evaluated the goodness of our pointwise prediction computing the RMSE for the polynomial model before mentioned. In this case we obtained $RMSE_{POLY} = 0.90$.

In both models we notice an increase of the R^2_{adj} with respect to their versions without dummies, as further confirmation that the inclusion of these variables is relevant to our analysis.
Moreover the two models present similar behavior in estimating the outgoing of our data, the Natural Splines setting seems to be better in terms of fitting since it has a slightly higher value for the R^2 and the R^2_{adj} parameters, but presents also a slightly higher value for the RMSE computed a posteriori, showing a worse prediction with respect to the Polynomial model. For the moment there is no model that can be labeled as "the best one", so we go on with our analysis keeping both the models.
Now let's discuss the role of the categorical variables in shifting the curves of the above presented models: by looking at the values of the regressors for the dummies we notice that the only two variables that have a positive shift on the curve are the *Marine Debris* and *Weekend* dummies, all the other contribute to a negative shift of the curve. In particular the highest possible shift is achieved when the expedition focuses on *Marine Debris* waste collection, performed during a *Weekend* in *Autumn*; meanwhile the lowest possible shift when the expedition focuses on *Underwater Cleanup*, during a *Weekday* in *Summer*.

4. Conformal Prediction

Finally we want to define Prediction Intervals for the expected number of collected items, given the number of Volunteers and the specific characteristics of the mission. In order to do so we used a Full Conformal approach, considering as Non Conformity Measure (NCM) the residuals of the model. We performed these estimates for both above mentioned models, that is using Polynomial and Natural Splines. Moreover we estimated the Conformal Prediction Intervals (CPI) for the two models in the cases with and without the categorical regressors. In the case without the categorical variables we assumed to have an equispaced grid of new data points covering the range of total volunteers recorded in 2018, and we performed the Conformal Prediction assuming as design matrix the corresponding transformations (Polynomial or Natural Splines) of those values. In the case including also the categorical variables, we expanded the design matrix adding some binary columns expressing the categorical regressors as dummy variables and for the prediction we performed a pointwise prediction assuming as new data the data recorded in 2018 (keeping together the observations). We report the obtained results in the following sections.

4.0.1 Polynomial Model

At first we considered the model expressing the relation between collected items and number of volunteers through a degree 8 polynomial transformation.

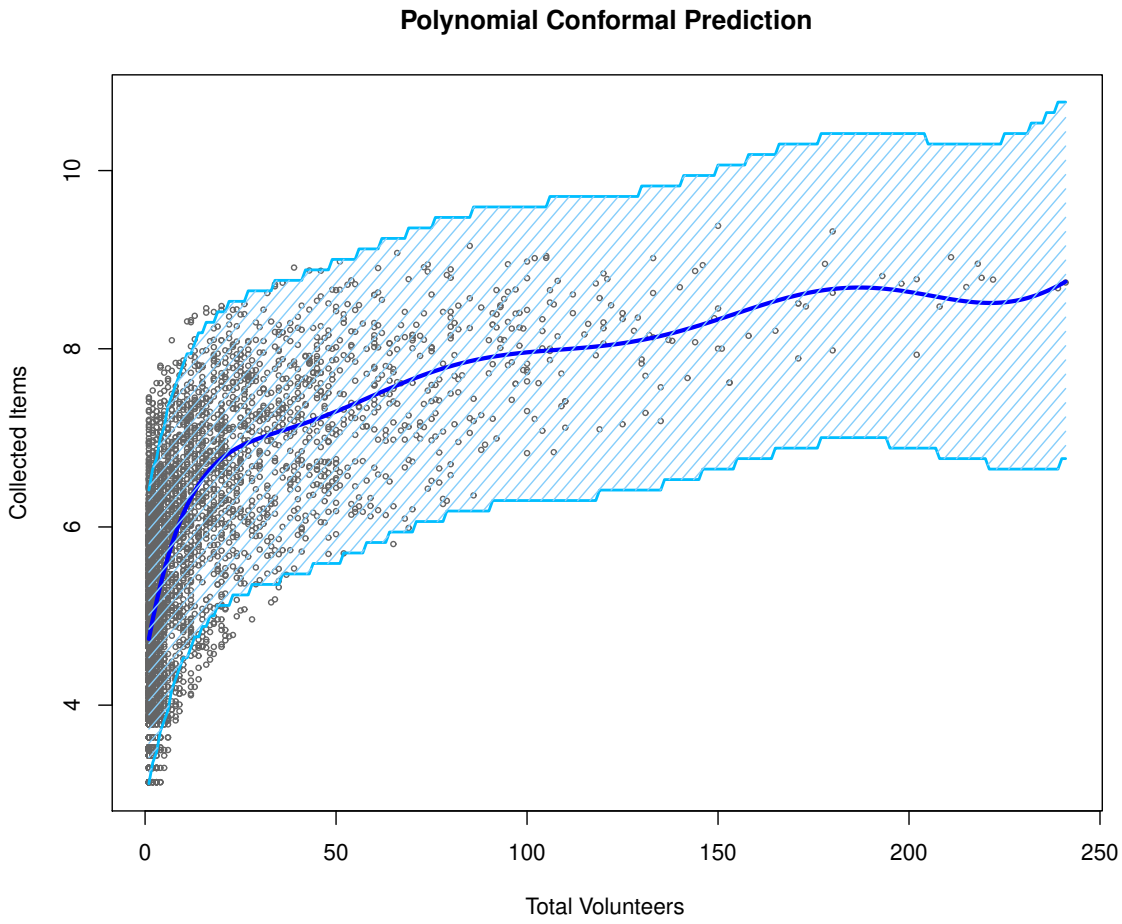

Figure 19: Polynomial CPI considering only the number of volunteers

Including also the categorical covariates for each observation, we obtained a specific predicted value for the expected number of collected items given not only the number of volunteers but also the peculiar characteristics of the planned mission. This more detailed formulation of the problem provided an upward or downward "shift" of the predicted data points, according to the value of the dummies characterizing the specific observation.

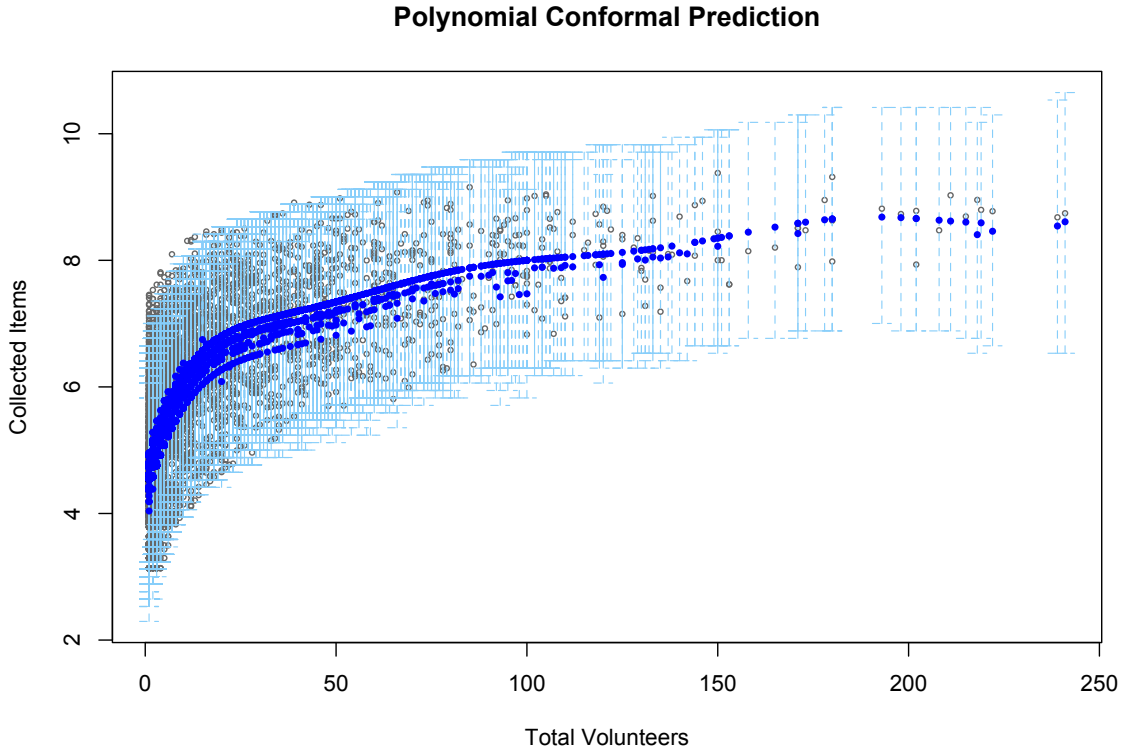

Figure 20: Polynomial CPI adding also categorical regressors

Considering this last more specific model we computed the proportion of true values (that is the real registered value of collected items during a specific mission) that were excluded from their corresponding conformal prediction intervals and we discovered that only the 5.65% of data didn't fit into the intervals.

4.0.2 Natural Splines Model

Also in the case in which we modeled the relation between numerical variables with a Natural Splines setting, we obtained a similar result in the two scenarios (with and without categorical regressors).

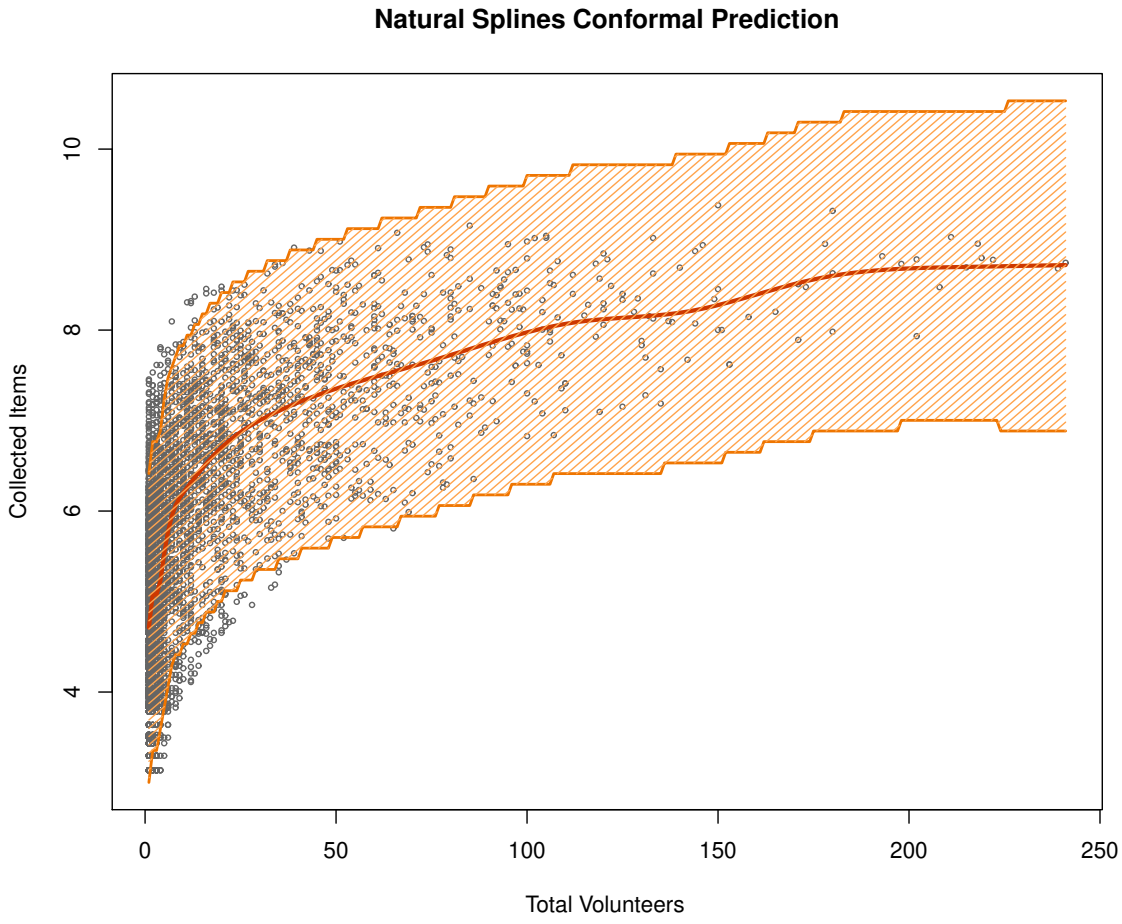

Figure 21: Natural Splines CPI considering only the number of volunteers

Also for this second model, after having included the categorical variables, we computed the proportion of data that weren't included in their corresponding conformal prediction intervals, discovering that also in this case only a few portion of test data didn't fit into our prediction set, more specifically the 5.7%.
At the end of our analysis the preferable model seems to be the one assuming the Natural Splines setting, considering the slightly better R^2 and R^2_{adj}, and the tighter prediction intervals.

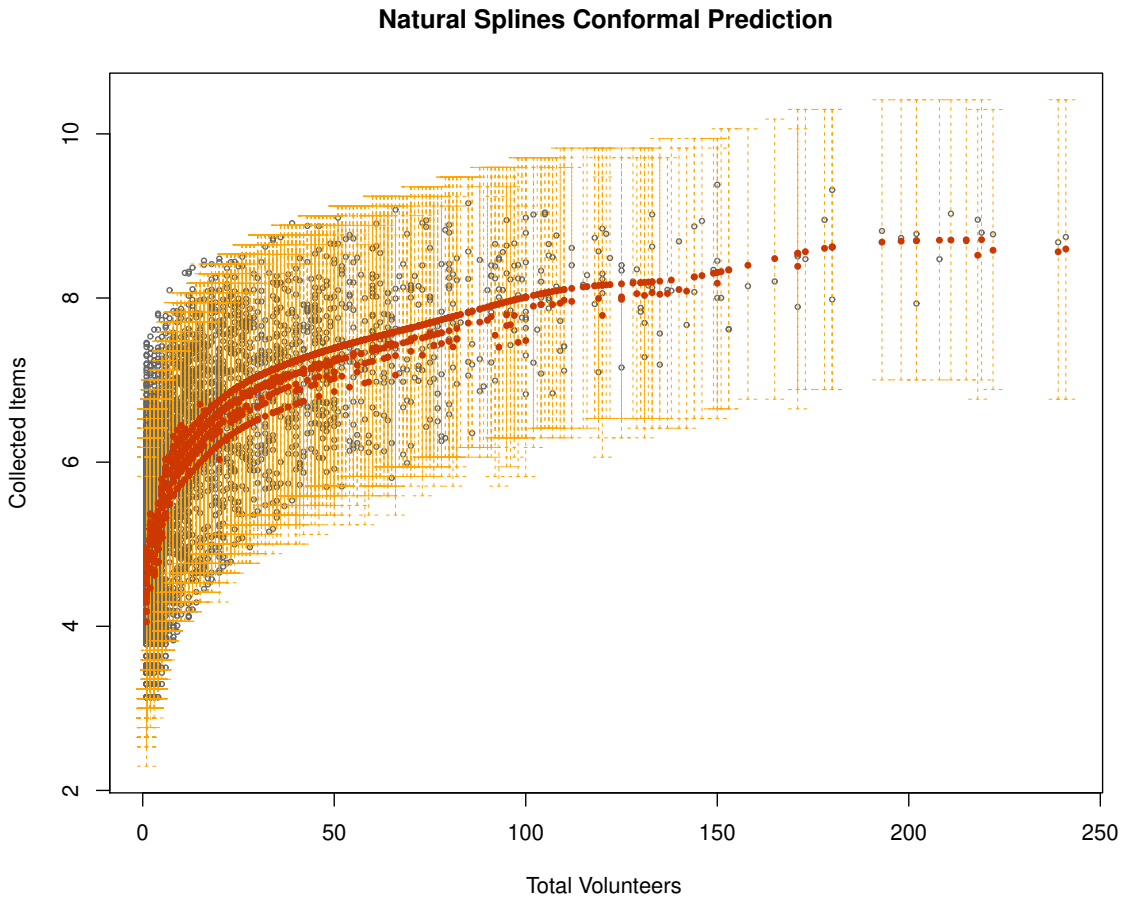

Figure 22: Natural Splines CPI adding also categorical regressors

5. Collected Plastic Prediction

After having stated a model for predicting the expected number of total collected items during a mission, we want to introduce a second model that aims at classifying the recyclable part of these items. In particular the classified items are divided into two main groups:

- Plastic and Foam
- Glass, Rubber, Lumber and Metal

These groups are defined among the classified items, i.e. among the items that the company was able to label with some recyclable material. In our dataset we have that during a mission the classified items are less or equal to the total collected items, this inequality is probably due to the presence of some non-classifiable items such as cigarettes, shoes, socks etc.
Our goal is to make inference about the amount of plastic items we expect to collect during a mission, given the total number of collected items.
The information we had about this classification was a percentage telling us how many items were labeled as plastic items among the totality, but in order to have a better fitting model we transformed this information obtaining the effective number of Plastic Items as follows:

$$PlasticItems = PlasticPercentage * ClassifiedItems$$

We then performed a logarithmic transformation of this quantity (always non-negative since represents the number of collected plastic items) in order to be coherent with the transformation of the number of total collected items. In this framework we see that the majority of the observations seem to follow a nearly linear trend, even if some data points are far away from the main data cloud.

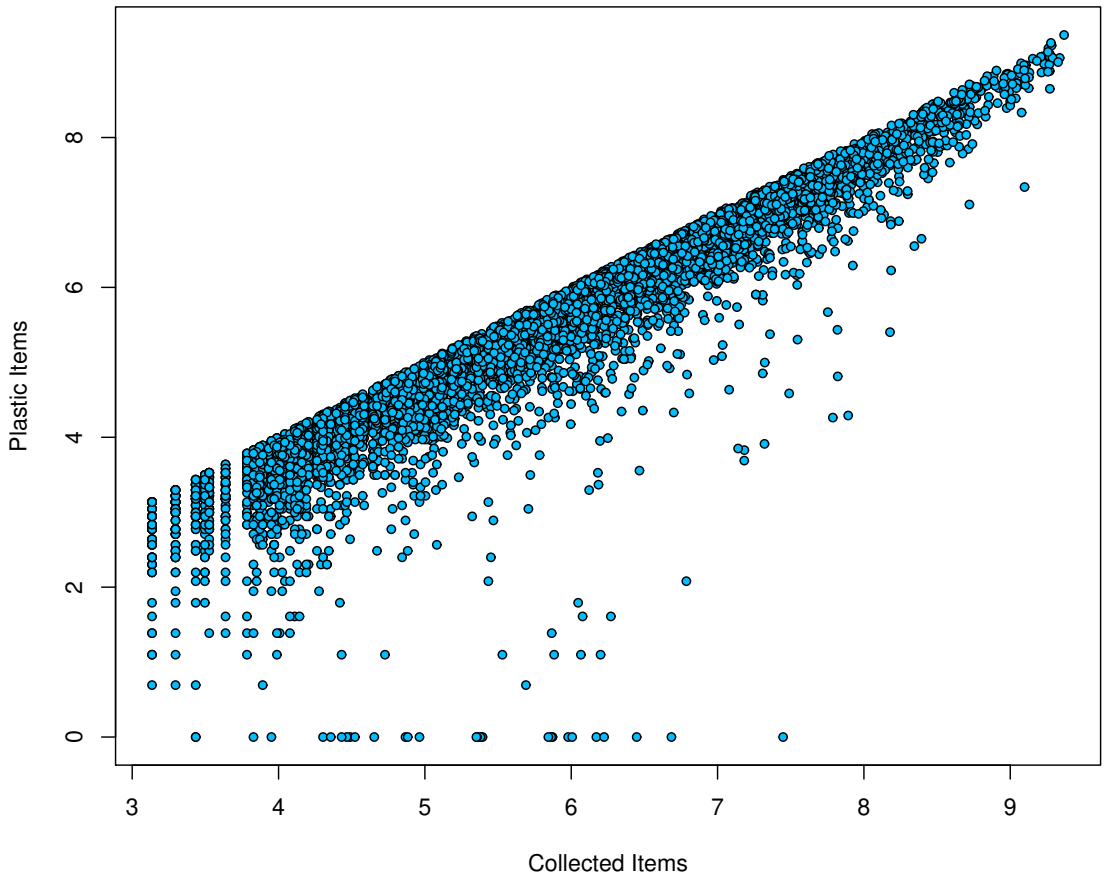

Figure 23: Collected plastic items versus total collected items distribution

Since this data weren't labeled as outliers in our first analysis (section 2.1) for the main model, we would like to avoid removing them from our observations. Moreover it's realistic that due to particular factors (local habits, specific places etc.) in some missions the amount of plastic discovered and collected by the volunteers may be lower than expected. For these reasons we decided to keep those "outlying" observations and, instead of discarding data, we adopted a robust approach to avoid estimates particularly affected by those data.
As can be seen in the following plot, the tolerance ellipse obtained through robust statistics is more skewed around the majority of data, while the classical one is heavily affected by the anomalous observations. The robust region was obtained using the *Minimum Covariance Determinant (MCD)* estimates for the mean value and the covariance matrix of our data. To compute these quantities we first need to define the following :

- fix $h \in \mathbb{N} : \lfloor \frac{n+p+1}{2} \leq h \leq n \rfloor$
- define $H \subset X$ s.t. $\#H = h$
- define $H^* := argmin_H det\big(\frac{1}{h}\sum_{x_i \in H}(x_i - \mu_H)(x_i - \mu_H)^T\big)$

At this point we can define the MCD robust estimates as:

$$\mu_{MCD} = \frac{1}{h}\sum_{x_i \in H^*} x_i \tag{1}$$

$$\Sigma_{MCD} = \frac{1}{h}\sum_{x_i \in H^*} (x_i - \mu_{MCD})(x_i - \mu_{MCD})^T \tag{2}$$

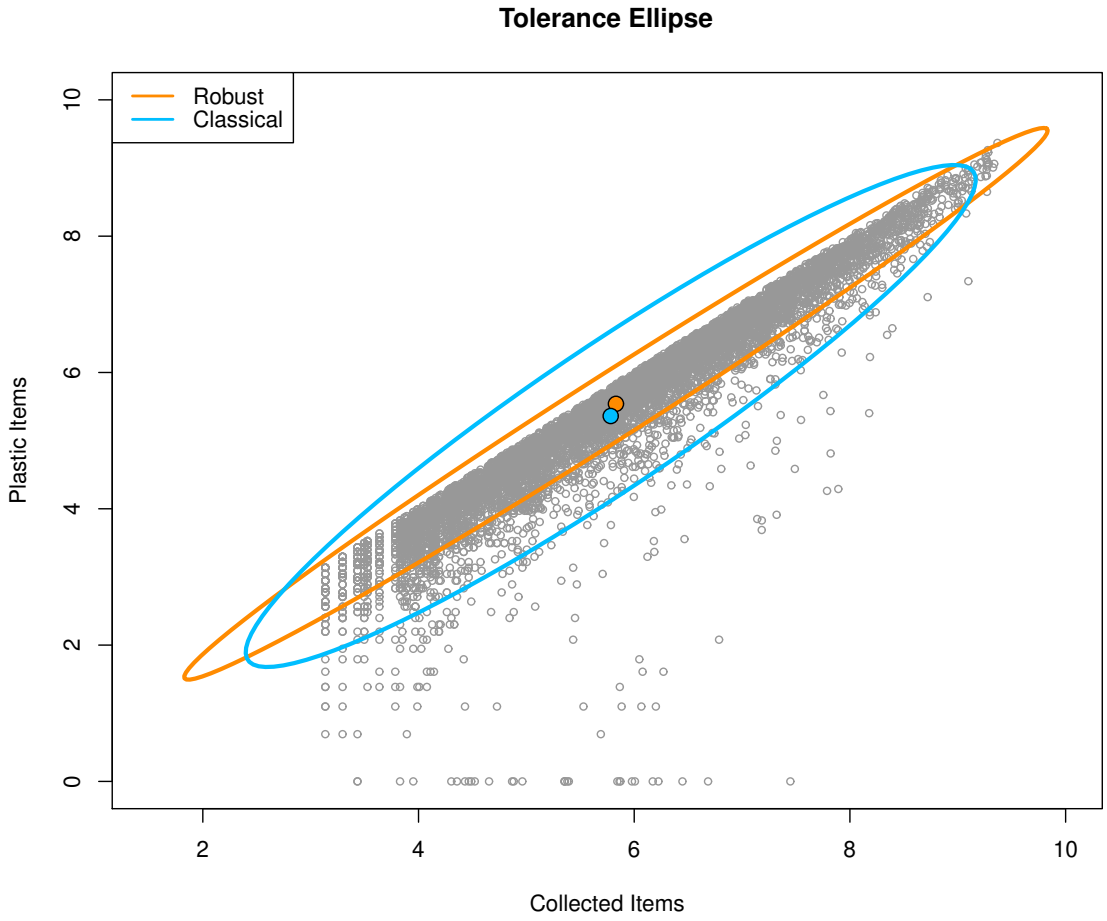

Figure 24: Tolerance Ellipse comparison

In this framework we want to fit a linear model for our data, that is:

$$PlasticItems = \beta_0 + \beta_1 TotalItems + \epsilon$$

To do so we considered many options :

- Classical OLS (Ordinary Least Squares)
  $\beta_{OLS} = argmin_\beta \sum_{i=1}^{n} r_i^2(\beta)$
- Robust LMS (Least Median of Squares)
  $\beta_{LMS} = argmin_\beta Me(r_i(\beta)^2)$
- Robust LTS (Least Trimmed Squares)
  $\beta_{LTS} = argmin_\beta \sum_{i=1}^{h} (r^2(\beta))_i$

And we compare the models performances in the figure below, looking at their goodness in interpolating the data cloud. It's easy to see that the regression line obtained using OLS is affected by the outlying observations that shifted the line to the bottom. The less affected line is the one obtained using the Least Median Squares method, as we could have expected.

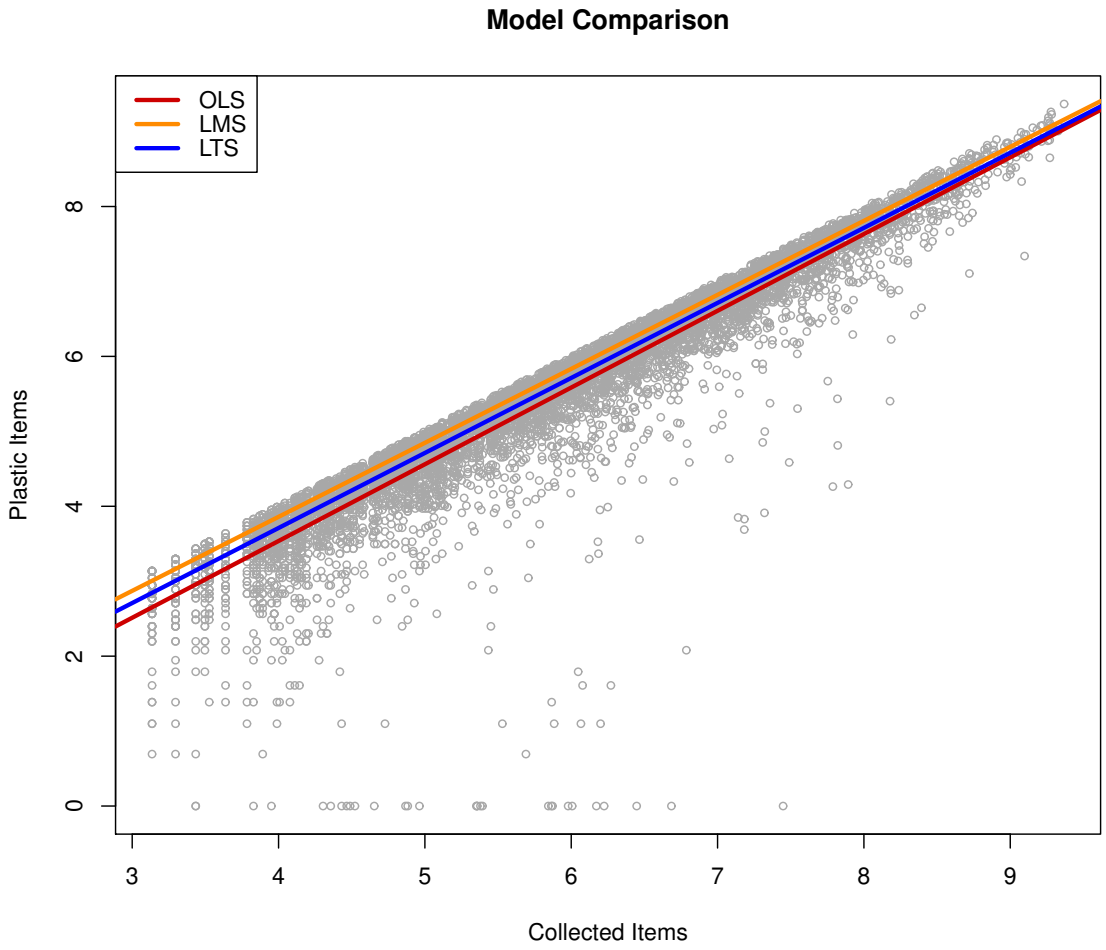

Figure 25: Regression Lines according to the considered models

We then wanted to try evaluating some confidence intervals for our estimated regression line. In order to do so we adopted a bootstrap approach, performing the bootstrap over the model residuals and looking for a confidence interval for the intercept of our regression line. This choice was due to the fact that the data cloud seems to have always the same trend, we just wanted to have a larger range over the data shifting upward and downward the regression line in order to include more observations. Since we are dealing with a lot of observations and we are assuming a very robust method, the estimate of this confidence interval was pretty tight assuming a level $\alpha = 0.05$ as reported in the plot. At the end we decided to not consider the computation of this intervals in the practical use of the model, since they are nearly overlapping the regression line.

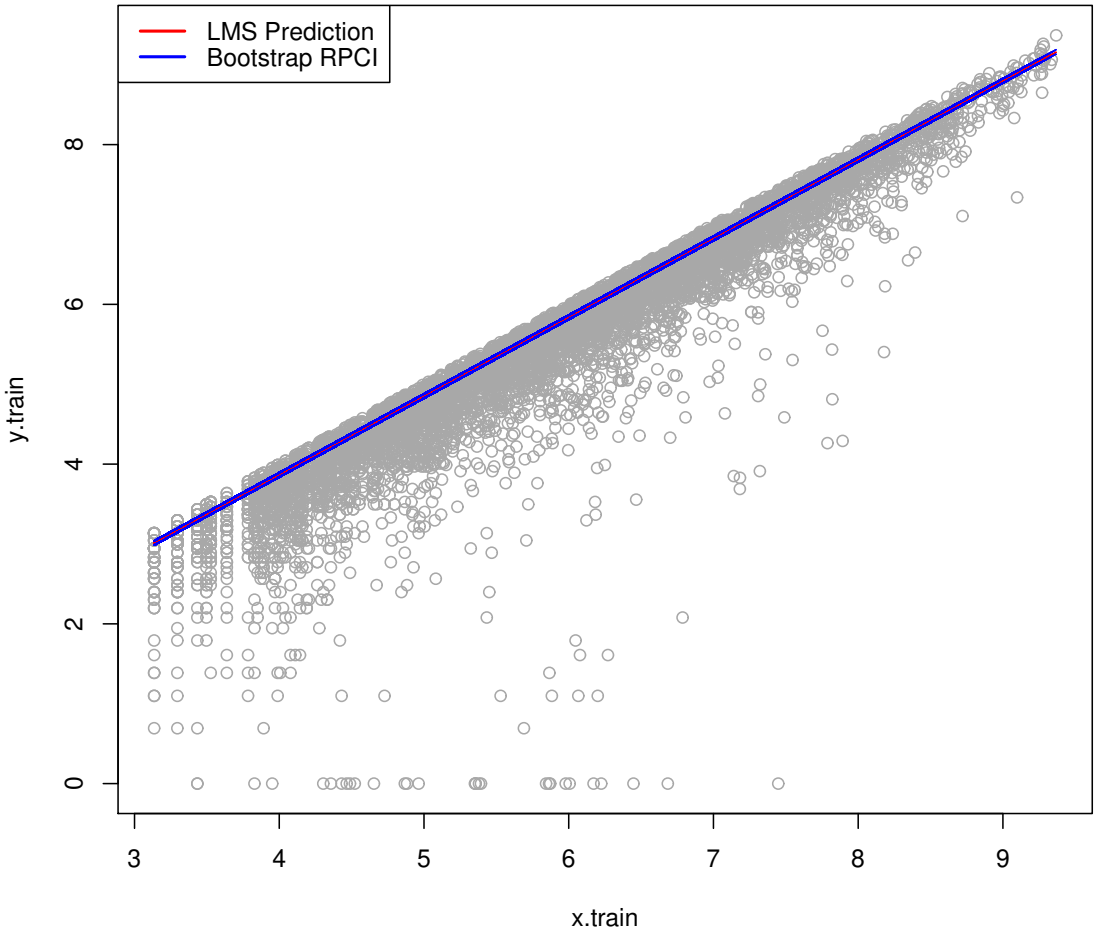

Figure 26: Regression Lines according to the considered models

6. A Realistic Example

To conclude our analysis we wanted to propose a real-life example. In order to do so, we randomly extracted one observation from the test dataset (2018) and assuming the corresponding features as the specific of a new mission, we wanted to check if the true value of collected items was actually into its corresponding conformal prediction interval and if it was close to the predicted value. The specific mission was defined as follows:

- Total Volunteers : 24
- Event Type : Land Cleanup
- Season : Autumn
- WE/WD : Weekend

Computing Conformal Prediction Intervals for the logarithm of the expected total collected items we obtain the following $CPI = [5.24, 8.53]$ with a prediction of 6.88, while the real value of the logarithm of collected items is equal to 6.35. We are pretty satisfied with this prediction and we want to use the predicted value 6.88 to make inference also about the expected number of plastic items. Adopting a LMS framework for the linear model of log plastic items we obtain an estimate of 6.71 while the true observed value was 6.05. Also in this case we are pretty satisfied with our estimate, also considering that for our purpose we don't need to have a perfect or precise estimate, we only need to have an idea of the amount of waste we will have to manage.

7. Appendix: Functional Analysis

At the beginning of our project we were going to do some functional analysis but we had to give up due to lack of data.
The initial dataset was composed of 38953 observations around the world:
most were related to 2018 (17384 observations), 12296 to 2017, 9273 to 2016 and only 1411 to 2015.
So we decided to only consider the years 2016,2017 and 2018 and we tried to create a new dataframe Countries/Months but we have seen that only 11 Countries out of 127 had at least 24 months out of 36 of observations, so too many Countries had too many months of missing values.
We also tried to consider only USA since it accounted for most of the observations (25188 observations out of 38953).

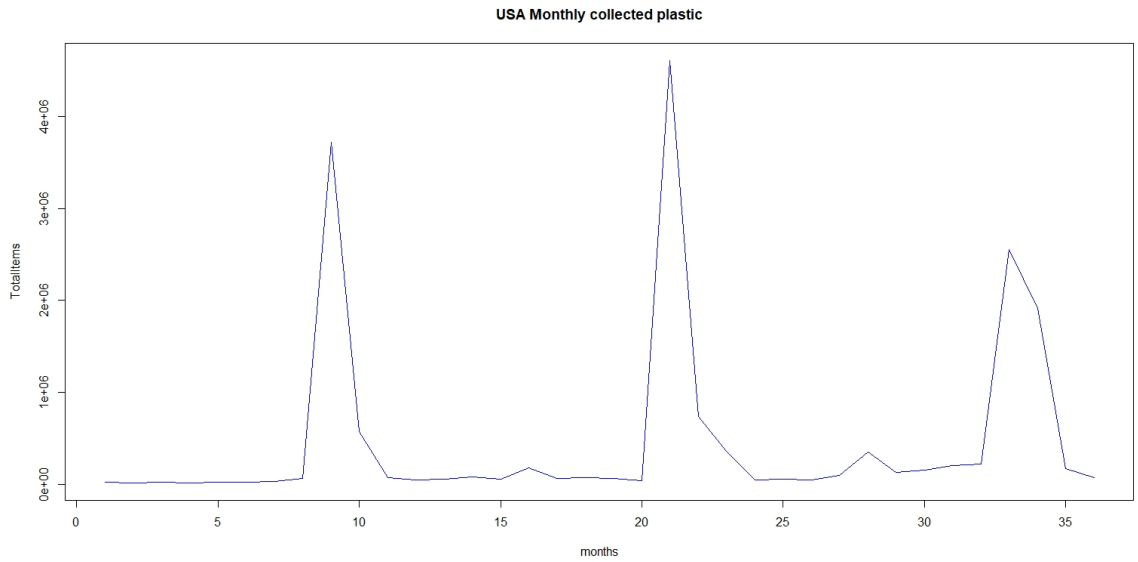

Figure 27: Functional Data: TotalItems vs Months of USA

From the plot above we can say that in USA the peaks occurred in the autumn months: September 2016, November 2017 and October/November 2018.
At that point we also tried to analyze more in depth the USA creating two new dataframe to confront daily and monthly expeditions of the 51 U.S. States.

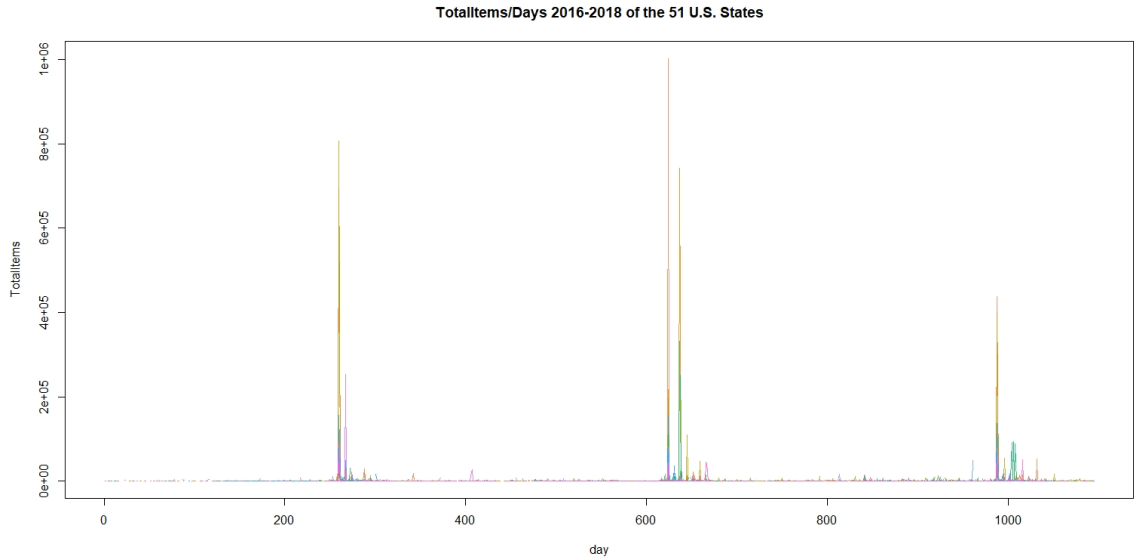

Figure 28: Functional Data: TotalItems vs Days of the 51 U.S. States

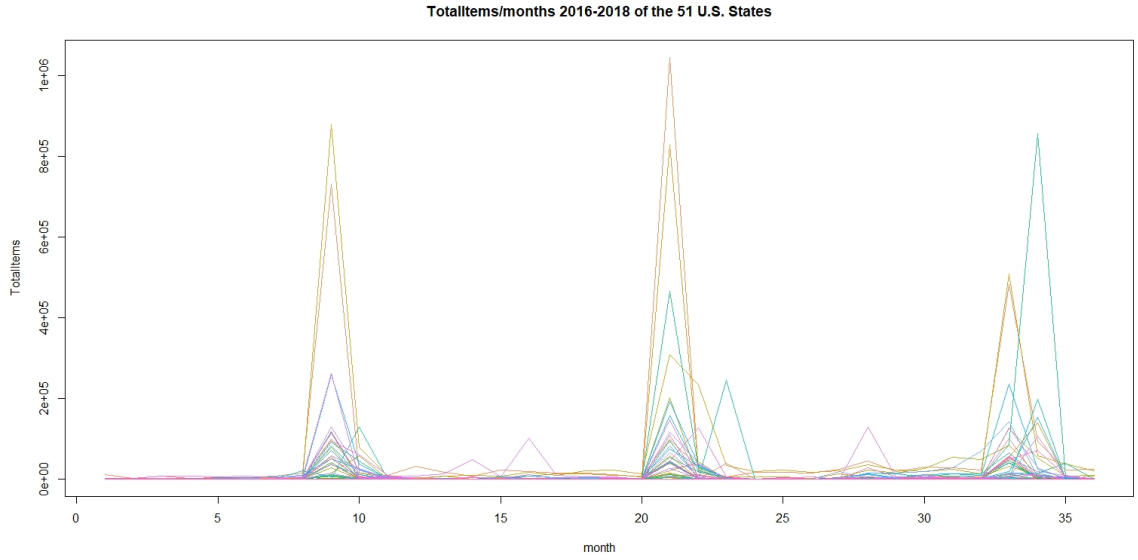

Figure 29: Functional Data: TotalItems vs Months of the 51 U.S. States

As we can see from the plots above there are too many missing values:

- 48796 missing values out of 55845 (51 states x 365 days x 3 years)
- 911 missing values out of 1836 (51 states x 36 months)

Therefore a functional analysis was discarded.