POLITECNICO
MILANO 1863

WASTE POLLUTION IN UNITED STATES OF AMERICA

Nonparametric Statistics – 27th of January 2022

Battistini Camilla, Frigeri Michela, Ronzulli Michael, Spina Pietro

THE WASTE MANAGEMENT PROBLEM

Can we know in advance how much waste will be collected in an expedition?

How can we better organize volunteers according to the type of expedition?

How much plastic do we expect to collect during an expedition? Can we optimize the recycling process?

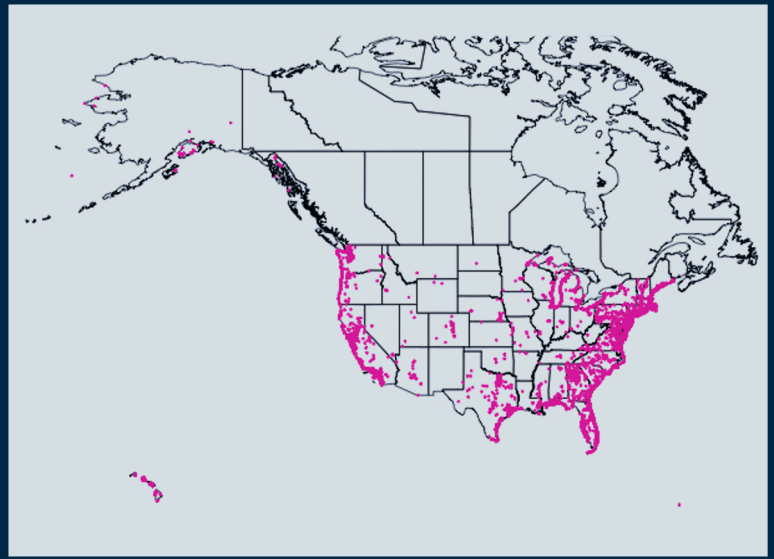
OUR DATASET

QUANTITATIVE DATA

- Total Item collected in the expedition
- Total number of Volunteers who took part in the expedition
- Percentage of items made of different materials

CATEGORICAL DATA

- Days of the week, months and years (Weekdays vs Weekend, Seasonality etc.)
- Event Type: Land Cleanup, Underwater Cleanup, Watercraft Cleanup and Marine Debris



GOALS OF OUR PROJECT

MODEL AMOUNT OF COLLECTED WASTE

To model the number of collected items during a certain event, according to the mission specific features in order to organize the waste disposal.

ROBUST FORECASTING

To predict the amount of these items belonging to the recycling categories of Plastic/Foam or Glass/Rubber/Lumber/Metal

A BRIEF RECAP

Data cleaning and outliers detection

Outliers detection through **Tukey depth** measure

Transformation of data

Log-transformation of number of items collected: better shape for model fitting

Covariates selection

From **ANOVA** significance test:

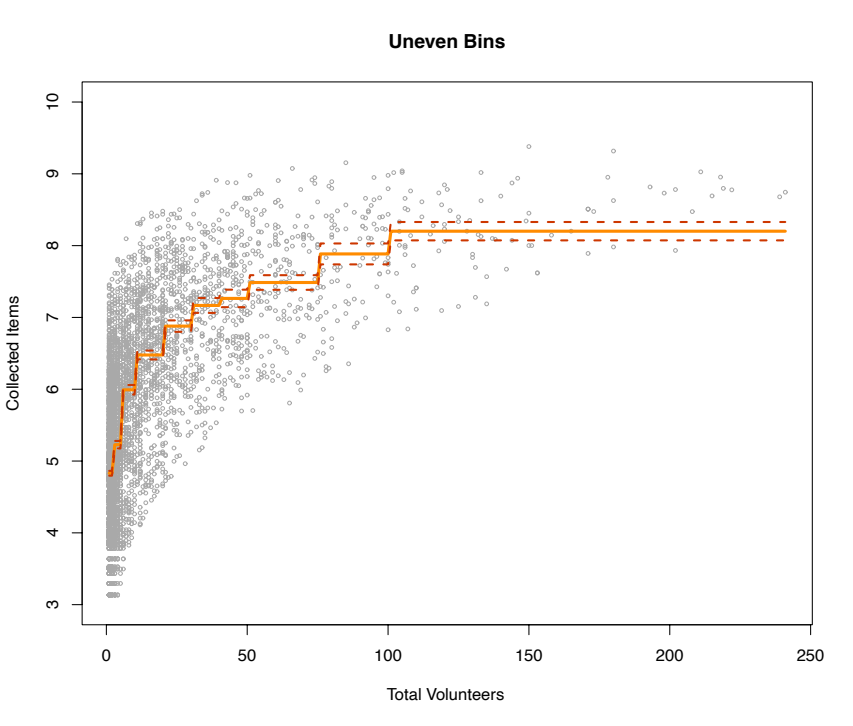
- Event type
- Weekend vs Weekday
- Seasonality
- Event type:Seasonality

Model proposal

- Step function regression with uneven bins
- Polynomial regression with degree up to 8
- Gaussian kernel regression
- Natural splines regression

MODEL CHOICE: step function regression

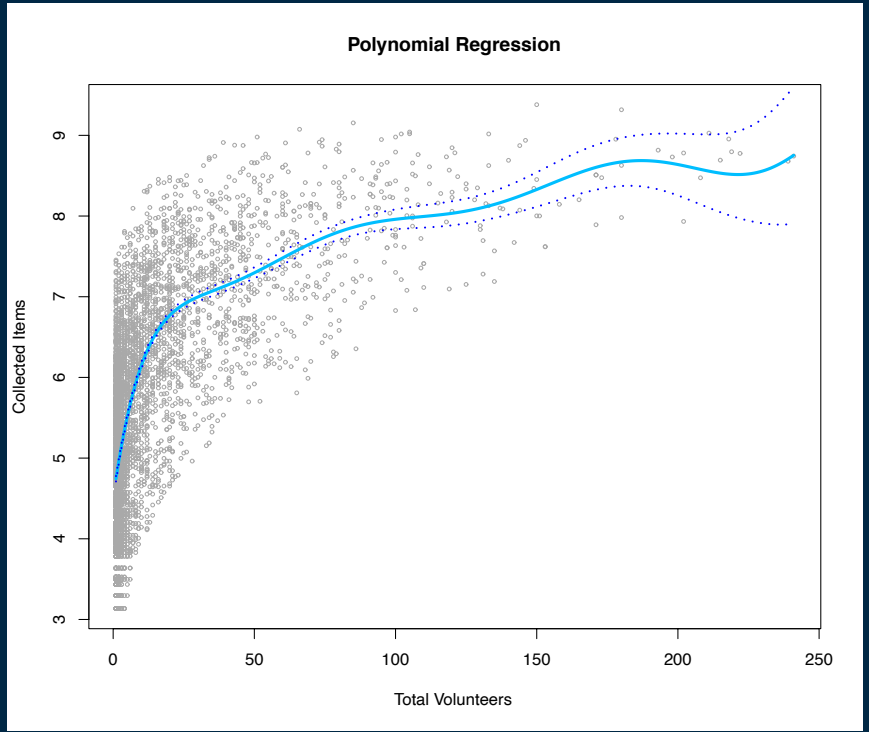
Goal: to build a regression model for log(TotalItems) ~ TotalVolunteers



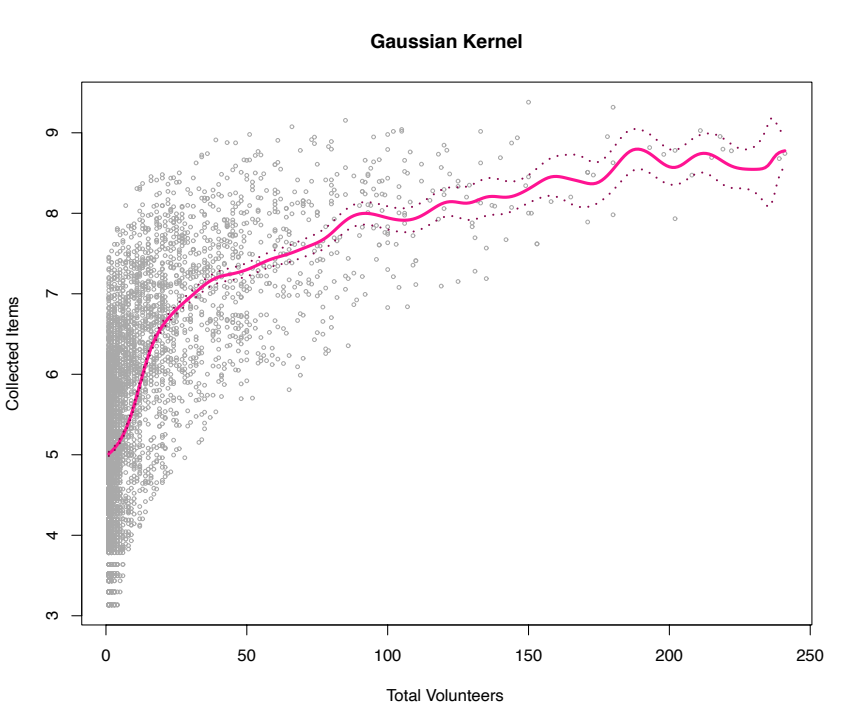
- We discard the choice of using even bins
- 10 uneven bins with an higher concentration towards small values of TotalVolunteers
- $R^2_{adj} = 0.559$
- Decently performing in terms of goodness of fit

MODEL CHOICE: polynomial regression

- Anova F-test to compare polynomial regression from degree 1 up to degree 10
- Polynomials are influential up to degree 8
- $R^2_{adj} = 0.568$
- One single very high leverage point



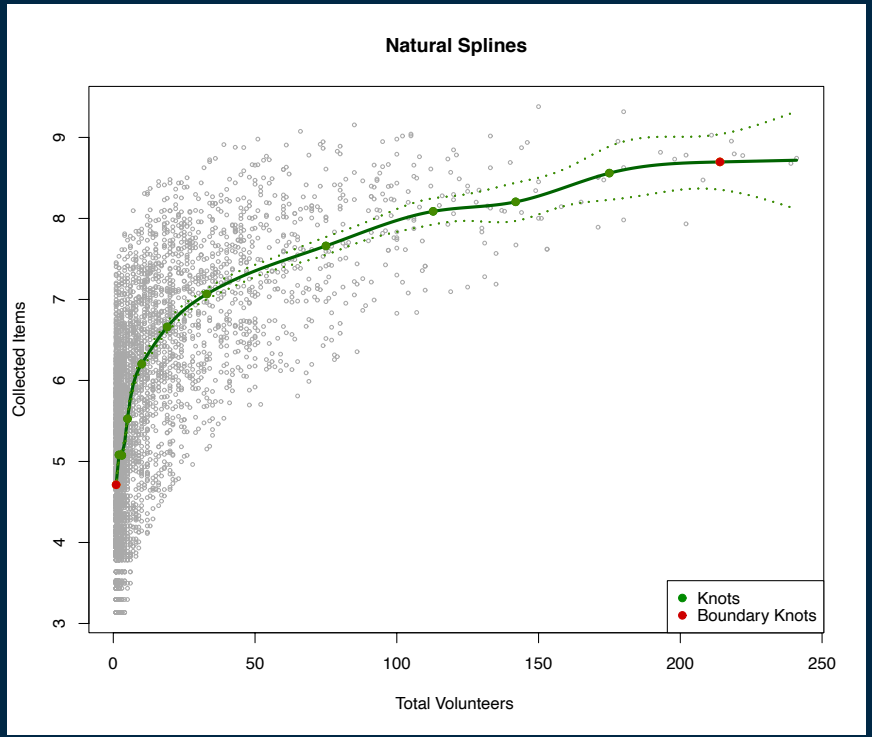
MODEL CHOICE: gaussian kernel regression



- We discard the choice of using Uniform kernel: really bad performing model due to lack of data in the bins
- Bandwidth set to 5: after different trials it seemed to produce a smooth curve while maintaining an increasing trend
- $R^2_{adj} = 0.532$

MODEL CHOICE: natural splines regression

- 12 knots set through a quantile driven approach
- Boundary knots at 1% and 99% quantiles
- $R^2_{adj} = 0.571$
- Good performance in terms of residuals



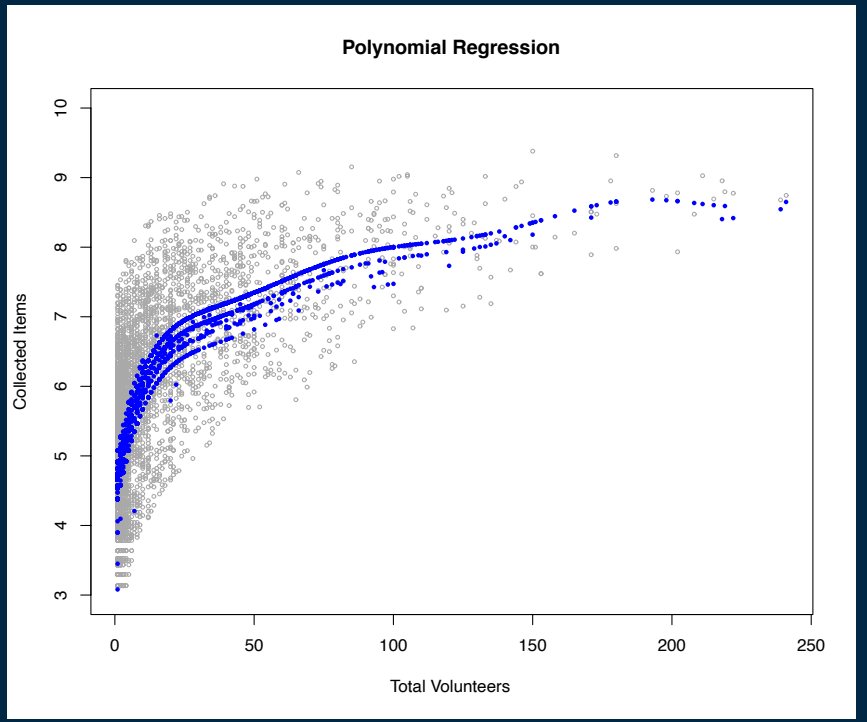
MODEL REFINEMENT: Categorical variables' contribute

Polinomial regression model

$$y = \beta_0 + \sum_{j=1}^{8} \beta_j x^j + \beta_9 W + \beta_{10} ET + \beta_{11} S + \beta_{12} ET{:}S + \varepsilon$$

- $R^2 = 0.5795$
- $R^2_{adj} = 0.5781$
- $RMSE = 0.90$

Improved fitting performances adding categorical regressors



MODEL REFINEMENT: Categorical variables' contribute

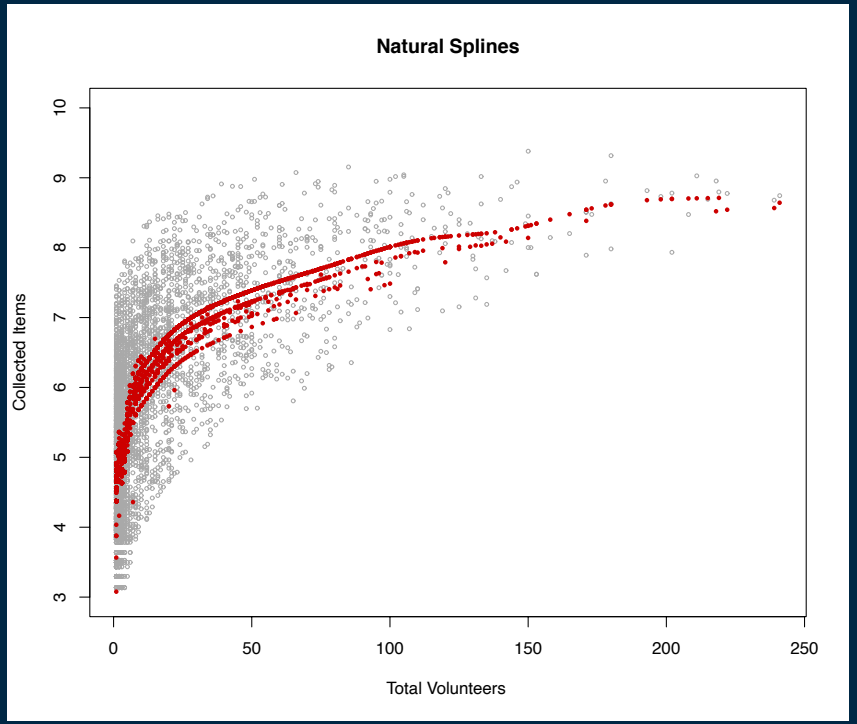
Natural splines regression model

$$y = \beta_0 + \sum_{j=1}^{3+m} \beta_j g_j(x) + \beta_{16} W + \beta_{17} ET + \beta_{18} S + \beta_{19} ET{:}S + \varepsilon$$

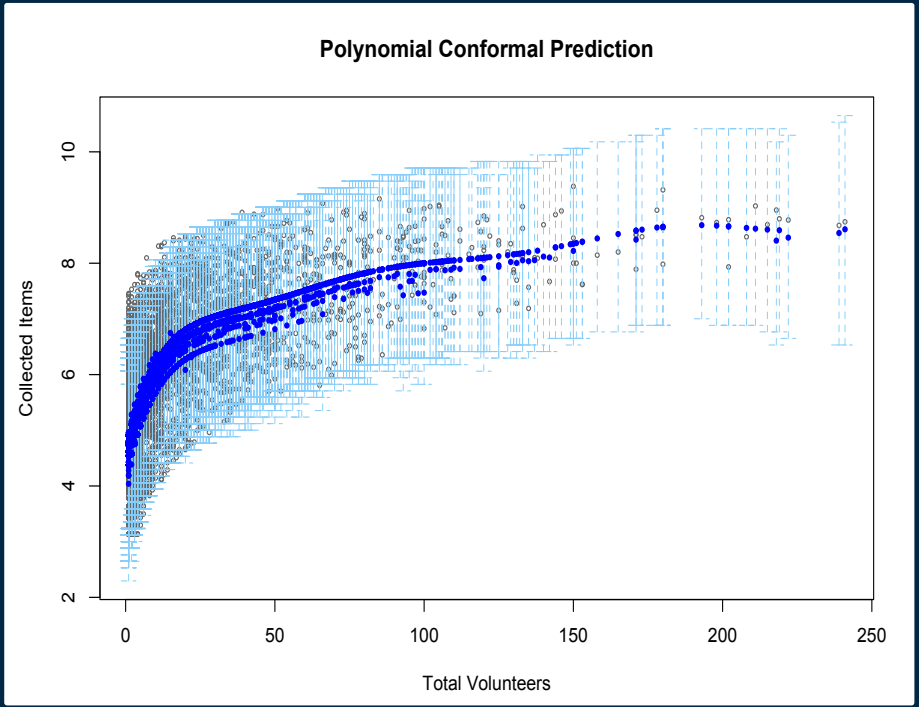
$m = 12$ nodes

- $R^2 = 0.582$
- $R^2_{adj} = 0.580$
- $RMSE = 0.91$

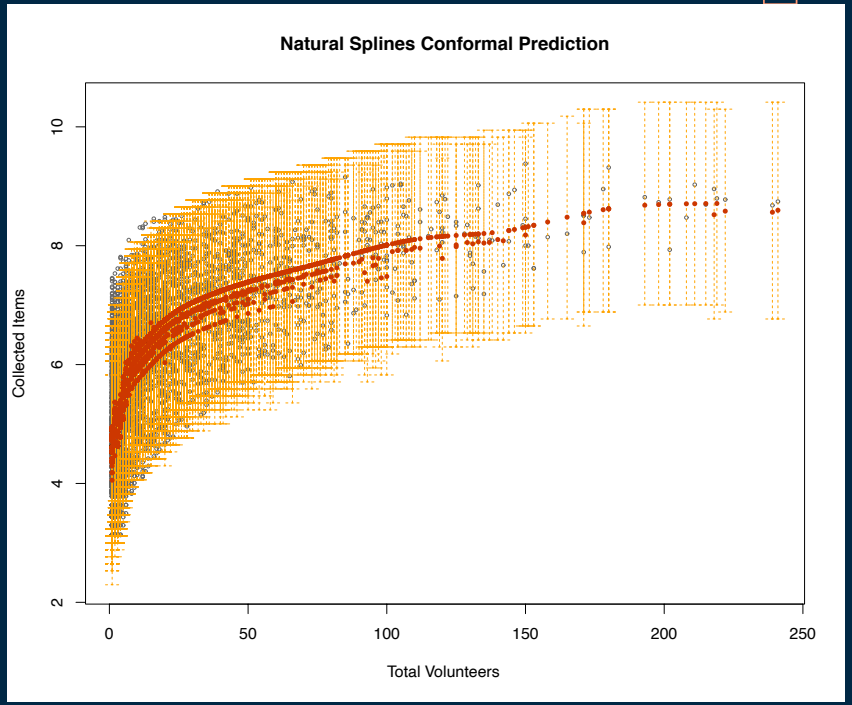
Improved fitting performances adding categorical regressors



CONFORMAL PREDICTION



Only the **5.65%** of data don't fit into the intervals.



Only the **5.70%** of data don’t fit into the intervals.

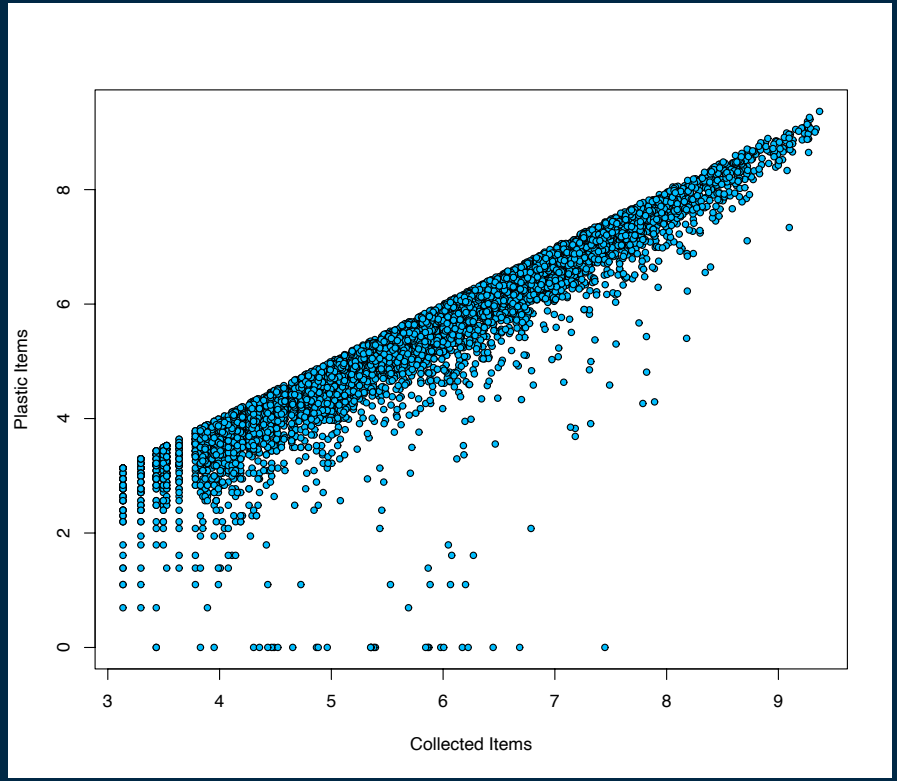
Collected Plastic Prediction

Items in our dataset are divided into:

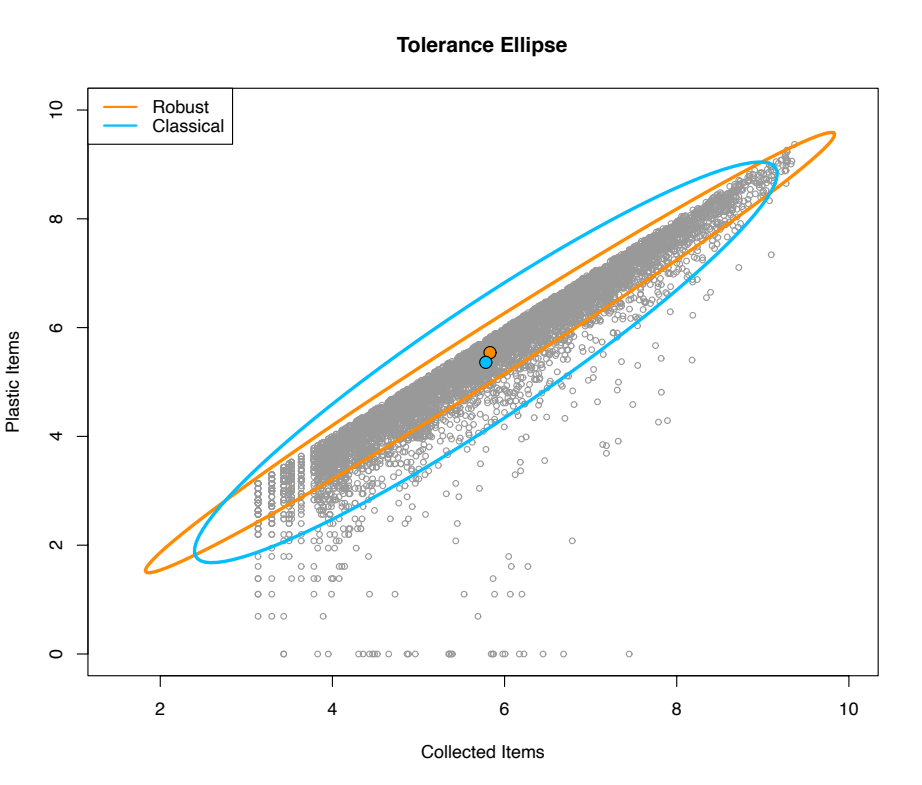
- Plastic and Foam
- Glass, Rubber, Lumber and Metal

We want to predict the amount of plastic items given the expected number of total collected items.

$$PlasticItems = PlasticPercentage * ClassifiedItems$$



Robust Statistics



Tolerance Ellipse (97.5%) comparison between:

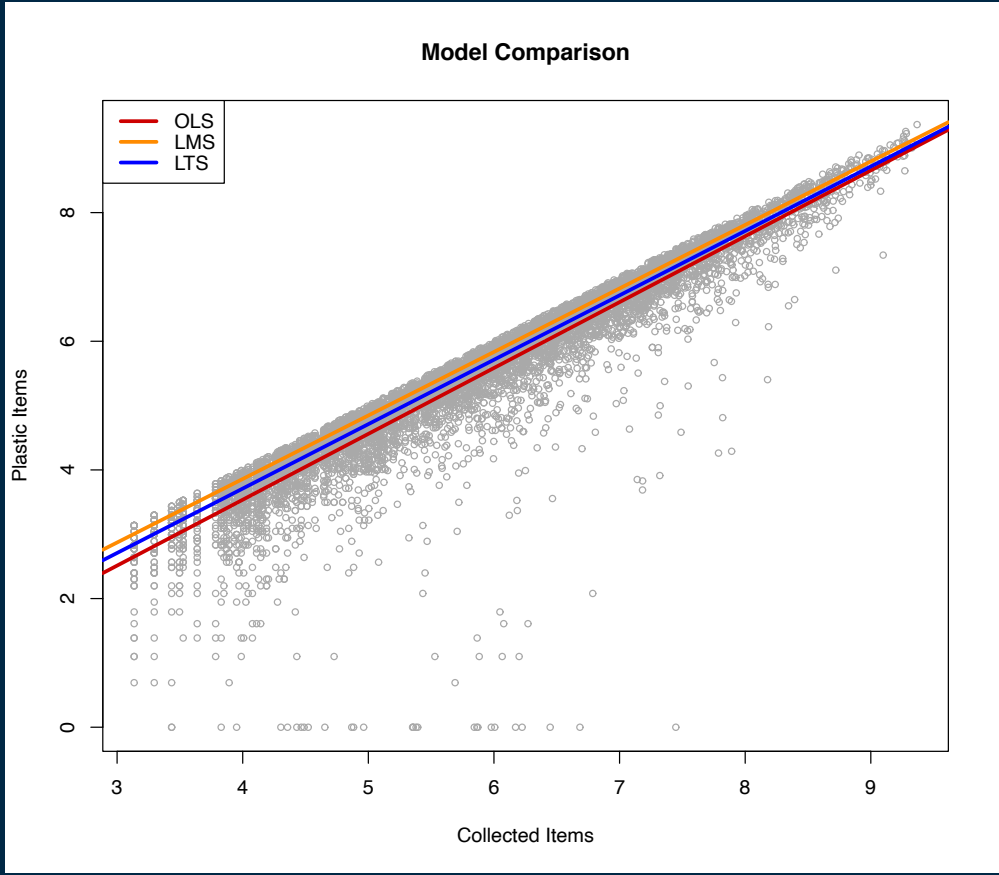
- the *classical region*
- the *robust region*
  obtained through **MCD** estimates for the mean value and the covariance matrix.

Linear Model

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

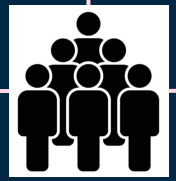
For the estimate of coefficients we considered three different models:

- Classical **OLS** (Ordinary Least Squares)
- Robust **LMS** (Least Median of Squares)
- Robust **LTS** (Least Trimmed Squares)



A REALISTIC EXAMPLE

24 Volunteers



Land Cleanup expedition

Scheduled in Autumn

During the weekend

Estimated LogItems: 6.88
Conformal Prediction Interval: [5.24, 8.53]

Real Items collected: 6.53

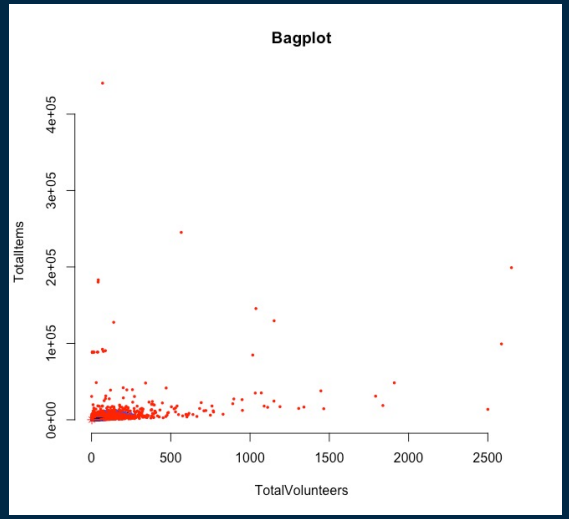
Estimated Log PlasticItems: 6.71

Real Log PlasticItems collected: 6.05

THANK YOU!

Do you have any questions?

OUTLIER DETECTION

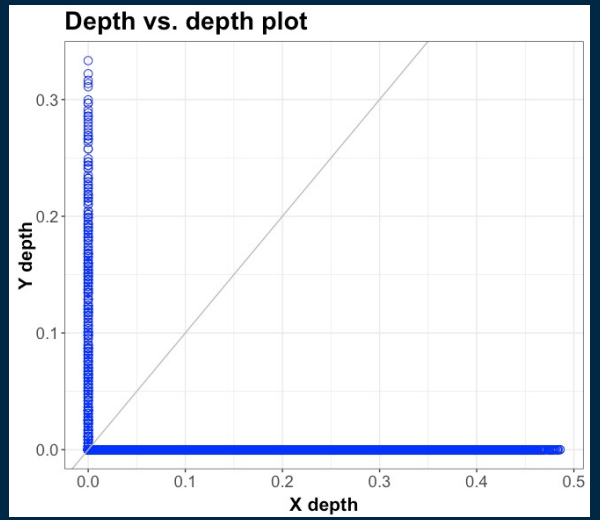


Bagplot

We identified many depth outliers relying on the bagplot (Tuckey Depth)

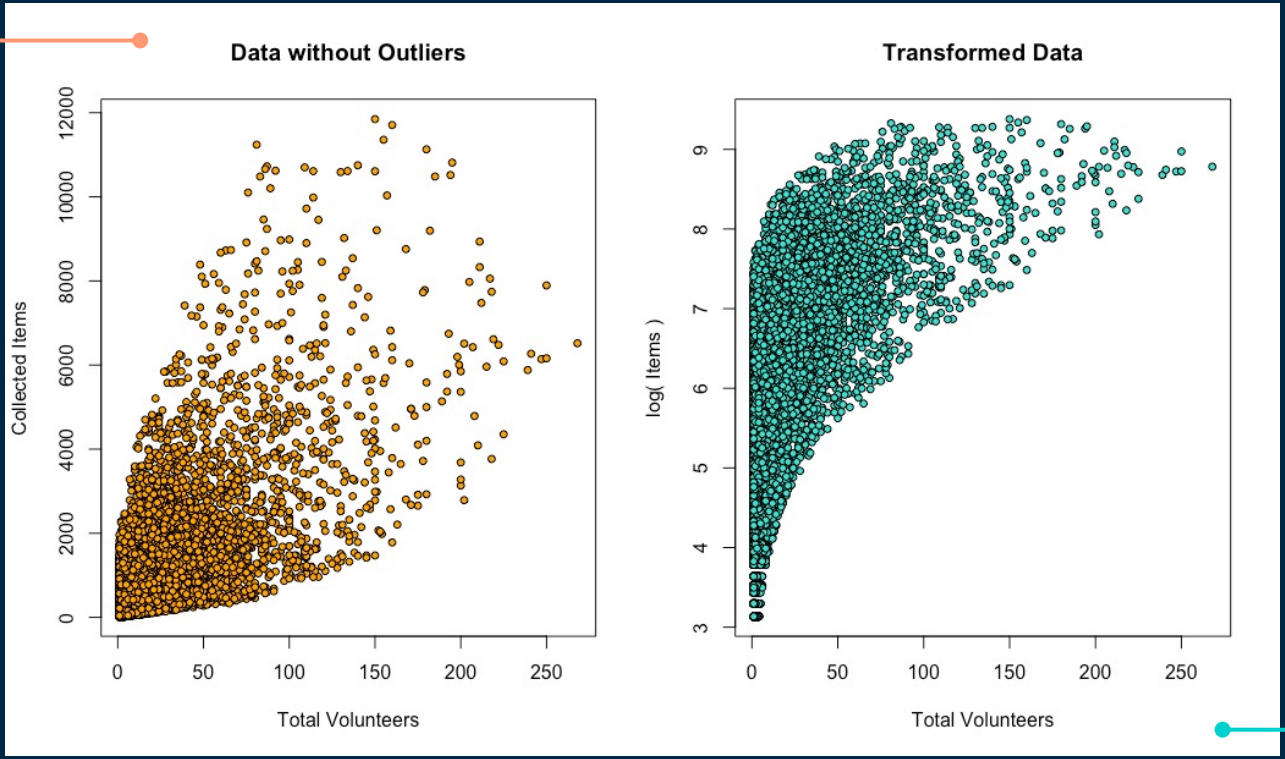
DD-Plot

Looking at the DDplot we checked the behavior of the outlying observations and we removed them from the dataset



DATA TRANSFORMATION

After removing the outliers, scatterrplot of volunteers against collected items appears as a cloud of data points



We then log-transformed the value of collected itemsin order to achieve a better shape for model fitting

ANOVA SIGNIFICANCE TEST

EventType

H_0: $\mu_0 = \mu_1 = \mu_2 = \mu_3$

vs

H_1: *at least one of the* μ_i *is different*

EventType significant

Weekend

H_0: $\mu_0 = \mu_1$

vs

H_1: $\mu_0 \neq \mu_1$

Weekend significant

Seasonality

H_0: $\mu_0 = \mu_1 = \mu_2 = \mu_3$

vs

H_1: *at least one of the* μ_i *is different*

Seasonality significant

EventType:seasonality

H_0: $\mu_0 = \mu_1 = \cdots = \mu_{16}$

vs

H_1: *at least one of the* μ_i *is different*

Interaction significant

Final model:

log(TotalItems) ~ TotalVolunteers + EventType + Weekend + Seasonality + EventType:Seasonality