Some applications of the bootstrap in Spanish Discards Sampling Scheme

Juan Santos, Aina Carbonell and Nélida Pérez

June 20, 2011

Abstract

Herein we propose a nested boostrap scheme to explore different sources of variability in discard data. The proposed tool resamples on trips, hauls within trips and lengthclasess within hauls to include all sources of variability identified in the Spanish DCR sampling program scheme. Megrim (*Lepidorhombus wiffiagonis*) and Hake (*Merluccius merluccius*) discard data collected in 2008 from `OTB-51` métier operating in ICES Division $VI - VII$, and Mediterranea `DEF`-Southern métier operating in the GSA 6S area were considered as case studies. Bootstrap Error and Coefficients of Variation associated to mean discards per trip were obtained and compared with asymptotic estimates. Furthermore, we use the bootstrap to quantify the contribution of the given sources of variability to global variability, and to assess simulated sampling scenarios differing each others in sampling effort. The last application of this method is for detecting outliers in 2008 sampled units (Trips). The good performance of the bootstrap method validates its use to obtain reliable error estimates in further regression and/or classification studies on discard data.

1 Introduction

Discarding is an unfair common fishing practice which impacts upon the health of fish populations and marine ecosystems (Rochet and Trenkel, 2005; Borges et al., 2005), and their importance is reflected in the fact that all EU countries are required under the EU Data Collection Regulation (DCR) to collect data on this catch fraction (CITE). As discards are considered to be high in some fleets (Alverson, 1994; Kelleher, 2006), estimations of this catch fraction are seen as an important source of fishing mortality information, otherwise hidden to stocks assessment or fisheries management. One common characteristic of discards is their high variability, conditioned by a wide range of factors, acting alone or interacting each others during the fishing process (Andrew and Pepperell, 1992; Alverson et al., 1994; Kennelly,20075; Rochet, 2002; Rochet and Trenkel, 2005). A challenge for researchers during the last decades have been how to address this feature inherent to discards in order to obtain quality estimates. Very cost sampling programs were enforced in European Countries in last decade achieving low precision estimates in general (Allen et al, 2002). Different methods were proposed for different authors to improve the quality of estimates , but doubts about the certainty of the estimates are still set out currently. As an example, the ICES Working Group of Hake, Monk and Megrim (WGHMM) have not carried out analytical assessments after 2006 for northern stocks of anglerfish, and the cause, among others, is the lack of confidence on discard information provided by the countries involved (ICES, 2010). It is clear therefore that more effort must be paid in understanding discard behaviour and measure the variance components aiming to optimize sampling schemes and sampling effort allocation. Herein we investigate the behavior of errors linked to discard estimations by using the resampling method known as bootstrap (Efron, 1979) on 2008 discard data obtained from two Spanish OTB mtiers; the mixed OTB-51, operating in ICES V-VII Divisions directed to Megrim (Lepidorhombus wiffiagonis), and the Mediterranean Demersal trawl fishery `DEF`-Southern directed to Hake (Merluccius merluccius) in GSA 6S areas. The bootstrap scheme used accounts for the multistage sampling design implemented within mtiers, by resampling between-within trip and within haul. Three applications of our bootstrap scheme is carried out in the paper:

- To use the bootstrap to investigate main sources of variability in the Spanish Sampling Program
- To simulate different sampling scenarios in order to assess how the precision of discard estimates could be improved
- To use of the bootstrap to identify outliers in sampling units and propose within metier post-stratification

2 Material and methods

2.1 Spanish Sampling design and raising procedures

The sampling strategy and raising procedures used in this paper is the standard implemented in the 'Spanish Discards Sampling Programme. (SDSP)', which is in accordance with the 'Workshop on Discard Sampling Methodology and Raising Procedures' guidelines (ICES, 2003).

The sampling design is based on the stratification of the fishing activities, dividing the fleet population into subpopulations by métier and quarter. Trip is considered the sampling unit to be sampled within mtier and the randomly of sampling depend on the fishers collaboration. In this regard, North Atlantic OTB-51 métier is being sampled randomly, while in Mediterranean `DEF`-Southern, sampling is carried out in a collaborative way with several ship owners. Sampling effort within trip (number of hauls) is approximately the 50% in a OTB-51 trip, being the sampled hauls selected sistematically by the observer. All hauls are sampled within a Mediterranean `DEF`-Southern trip.

For a given sampled métier, sampling on discarded species is carried out as follows:

Let h_{ij} be the j-th ($j = 1, \ldots, J$) sampled haul in sampled trip i ($i = 1, \ldots, t$), and d^s_{ij} be a randow sample drawn from the discarded catch d_{ij} from h_{ij}.Let

$$r_{ij} = \frac{d^s_{ij}}{d_{ij}} \tag{1}$$

be the ratio of the sampled weigth to the total weight of discards.

Let f_{ijlk} be the k-th (k= 1,..., n) fish of size l sampled in d^s_{ij}, and $F_{ijl} = \sum_{k=1}^{n} f_{ijlk}$. Biomass by size can be obtained using the species weigth-length relationship available for the species:

$$F^w_{ijl} = \sum_{k=1}^{n} f_{ijlk} \times a \times l^b \tag{2}$$

2.1.1 Trip level

Let

$$y_{ijl} = F_{ijl} \times r_{ijl} \tag{3}$$

be the estimated numbers of individuals of size l discarded in haul j and,

$$y^w_{ijl} = F^w_{ijl} \times r_{ij} \tag{4}$$

the estimated discards in terms of biomass. the mean discards for size l in trip i can be calculated as follows,

$$\bar{y}_{il} = \frac{1}{J} \sum_{j=1}^{J} y_{ijl} \tag{5}$$

with variance

$$\sigma^2_{\bar{y}_{il}} = \frac{1}{J-1} \sum_{j=1}^{J} (y_{ijl} - \bar{y}_{il})^2 \tag{6}$$

if J is the total number of hauls carried out in trip i , the estimated total discards in numbers by size is:

$$Y_i = \sum_{j=1}^{J} y_{ijl} \tag{7}$$

else,

$$Y_i = \bar{y}_{il} \times H_i \tag{8}$$

with H_i being the total number of hauls (sampled + unsampled). The variance associated to (8) is

$$\sigma_{Y_i} = (1 - \frac{J}{H}) \times H^2 \times \frac{Var(\bar{y}_{il})}{J} \tag{9}$$

2.1.2 strata level

- Raising by number of trips (assumed known)

Mean discarded by trip is estimated to be

$$\bar{Y} = \frac{1}{t} \sum_{i=1}^{t} \times Y_i \tag{10}$$

with associated variance

$$\sigma_{\bar{Y}} = \frac{1}{t-1} \sum_{i=1}^{t} (Y_i - \bar{Y})^2 \tag{11}$$

(10) and (11) can be raised to the total fishing effort of the fleet (T), to obtain a estimation of total Discarded (D) of the fleet:

$$D = \bar{Y} \times T \tag{12}$$

with variance

$$Var(D) = (1 - \frac{t}{T}) \times T^2 \times \frac{Var(\bar{Y})}{t} \tag{13}$$

2.2 Bootstrap scheme

Classic estimations of a given characteristic θ of a variable of interest X are based on an asymptotic approach, where θ is estimated by $\hat{\theta}(X_1, \ldots, X_n)$. This asymptotic approach may display certain limitations in practice, since their performance will depend on the information furnished about the population by the sample. Hence, when not enough data are available, a good approximation of the distribution of θ will not be obtained, and so the asymptotic method will not produce good results. The bootstrap enable the sample distribution of the target statistic to be obtained by simulating a high number of random samples directly constructed on the basis of initially observed data. The simplest version of the bootstrap (naive bootstrap), is implemented in the Spanish Discard Sampling Scheme in order to obtain bootstrap discard statistic at métier level, in accordance with the following procedure:

1. A random trip pseudosample is artificially created by resampling with replacement on sampled trips $t = t_1, \ldots, t_n$. In other words, after the extraction of an trip, this is replaced in the original sample such that it can be chosen again.

2. hauls and lengthclassess within hauls are subsequently resampled within each $t*_i$ from the previous step, keeping the original number of hauls and individuals sampled.

3. Estimation of mean discards by trip $\bar{Y}^*$ is obtained by using the protocols from the original discard scheme.

4. Steps 1 , 2 and 3 are repeated a large number (B) of times, so as to obtain bootstrap values $Y_1^*, \ldots, Y_b^*$. Finally, the distribution of Y and its corresponding quantiles is approximated by means of the values $Y_1^*, \ldots, Y_b^*$, without any assumptions having been made as to the theoretical distribution to which the latter conforms.

Bootstrap mean discarded per-trip and associated bootstrap errors are compared with Asymptotic results in first instance. The bootstrap scheme is further used in different simulations designed to investigate ways of optimizing sampling effort and to evaluate effects of each source of errors (Trips, hauls and the length sampling variability) in the uncertainty of mean discarded per trip.

2.3 Measures for comparison

To compare the results from the boostrap simulations and the asymptotic results, the following statistics were calculated:

- $\sigma^*_{\bar{Y}_l} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\bar{Y}^*_{b,l} - \bar{\bar{Y}}_l^*)^2}$
- $CV^*_{\bar{Y}_l} = 100 \times \frac{\sigma_{\bar{Y}_l^*}}{\bar{Y}_l^*}$
- $ECVB_l = \frac{1}{\bar{Y}_l}\sqrt{\frac{1}{B}\sum_{b=1}^{B}(\bar{Y}_l^* - \bar{Y}_l)^2}$

The ECVB measures precision and bias together, as this statistics measures root mean square diferences between bootstrap mean discarded per trip ($\bar{Y}_l^*$) and the deterministic value ($\bar{Y}_l$), instead the natural comparation between $\bar{Y}_l^*$ and $\bar{\bar{Y}}_l^*$.

2.4 Simulations

2.4.1 Sources of variation study

To evaluate the error due to each one of the potential sources of variability (Trips, hauls and within haul sampling), three different bootstrap simulations were carried out using the OTB-51 data, leaving-one-out of the sources of variation in each of the simulations:

- SIM_t : Bootstrapping hauls and sampled individuals
- SIM_h : Bootstrapping Trips and sampled individuals
- SIM_w : Bootstrapping Trips and hauls

Which can be compared with:

- SIM_{ALL}: Original Bootstrap scheme including all sources of variability

The underlying idea is that when using bootstrap leaving-one-out of the sources of error, it is expected a decrease in the error associated with mean discarded by trip estimation. Thus, we can assess the importance of each of the included sources of variability by comparing errors from each of the simulations with SIM_{ALL}.

2.4.2 Sampling design study

In a second stage, The bootstrap is used to investigate ways of optimizing sampling effort, by simulating different sampling scenarios differing each others in number of sampled trips and number of sampled hauls within trips.In this regard, second stage is only used for the `OTB-51` data, as 100% of the hauls are already sampled by the observers. Results from this simulations are compared by using their respectives CV and $ECVB$.

2.5 Outlier detection

2.5.1 Approach 1: Linear approximation for outlier assessment.

Finally, we use our bootstrap scheme to detect outliers over the trips sampled during the sampling Mediterranean sampling year, by using regression methods in the linear approximation of the effect of each sampled trip in the error associated to discard estimation.This linear approximation procedure was obtained by Efron (1990).

Let $M_{b,i}$ be the number of times the original trip t_i is included in the b^{th} bootstrap sample and let $P_{b,i}^{=}\frac{M_{b,i}}{n}$. A linear regression without intercept of the form

$$\theta_b^* = sum_{i=1}^{n}\beta_i P_{b,i}^* + \epsilon^* \tag{14}$$

being θ_b^* the bootstrap error associated to the b^{th} estimation of mean discarded at trip level. The regression yields β_i coefficients which are centered $L_i = \beta_i - \beta$ to obtain a linear approximation of the effect of each of the sampled trips in the error of estimations used in this paper to detect those trips with extreme influence that could be considered as outliers.

2.5.2 Approach 2: Jackniffe-after-boostrap and bootstrap hypothesis testing.

3 Results

3.1 Sampling effort

A total of 7 trips were sampled randomly from the OTB-51 during 2008 sampling year. In average, $\sim$ 71 of valid hauls per trip (H) were carried out , being sampled $\sim$ 35 (J) in average representing $\sim$ 50% (p) of sampling coverage. Discarded Megrim were found in all trips, being sampled $\sim$ 287 fishes within d^s_{ij} per sampled haul. Length size of the sampled individuals ranges $80 - 280$. Scarce Hake discards were found in the mediterranean métier, being recorded in only 3 trips for a total of 14. Total hauls set in the sampling trips sange from 2 to 4, being all sampled by the observer.

	J	H	p	mean.r_i	ind.sampled	mean.size	range.sizes
Trip 1	33	76	0.43	0.01	447.00	198.00	80-280
Trip 2	35	74	0.47	0.03	609.00	190.00	80-280
Trip 3	38	78	0.49	0.02	290.00	199.00	80-280
Trip 4	35	53	0.66	0.01	102.00	174.00	80-270
Trip 5	41	80	0.51	0.01	136.00	177.00	80-280
Trip 6	32	64	0.50	0.01	215.00	181.00	80-270
Trip 7	33	70	0.47	0.02	213.00	202.00	80-280

Table 1: Trips sampled in OTB-51 during 2008 Discard Sampling Program.

	J	H	p	mean.r_i	ind.sampled	mean.size	range.sizes
Trip.1	2	2	1.00	0.0042	5	150	150-150
Trip.2	2	2	1.00	0.0048	3	180	180-180
Trip.3	2	2	1.00	0	0	NaN	-
Trip.4	3	3	1.00	0	0	NaN	-
Trip.5	2	2	1.00	1	162	143	90-230
Trip.6	3	3	1.00	0	0	NaN	-
Trip.7	3	3	1.00	0	0	NaN	-
Trip.8	2	2	1.00	0	0	NaN	-
Trip.9	3	3	1.00	0	0	NaN	-
Trip.10	2	2	1.00	0	0	NaN	-
Trip.11	3	3	1.00	0	0	NaN	-
Trip.12	2	2	1.00	0	0	NaN	-
Trip.13	4	4	1.00	-	-	-	-
Trip.14	2	2	1.00	-	-	-	-

Table 2: Trips sampled in DEF-Southern during 2008 Discard Sampling Program.

3.2 Estimations in real sampling scenary

Table **??** summarize original and bootstrap discard estimations for the métiers under study. Results indicates that the bootstrap scheme yield higher error and lower discard amounts in all estimations. Figures 1 and 2 show original estimates togheter with the bootstrap estimates by length classes. Consistent with the information from table **??** error values obtained from the bootstrap are in general higher than asymptotic errors. The difference in case of the atlantic mtier is lower than in the mediterranean, where both abundance and associated error clearly unmatch asymptotic estimation, and this behaviour is mostly observed for length sizes with higher discards. "CV is used in Figures **??** and **??** to compare the errors between lengthclassess. There is a clear negative relation between CV and discard abundance.

	n.medio	n.error	w.medio	w.error
asymptotic	48562.40	4692.05	2102.88	187.86
bootstrap	46434.99	5396.53	2013.39	227.67
Change(%)	-4.38	15.01	-4.26	21.19

Table 3: Asymptotic and bootstrap OTB-51 Megrim discards per trip and associated errors in year 2008.Results aggregated omiting length information

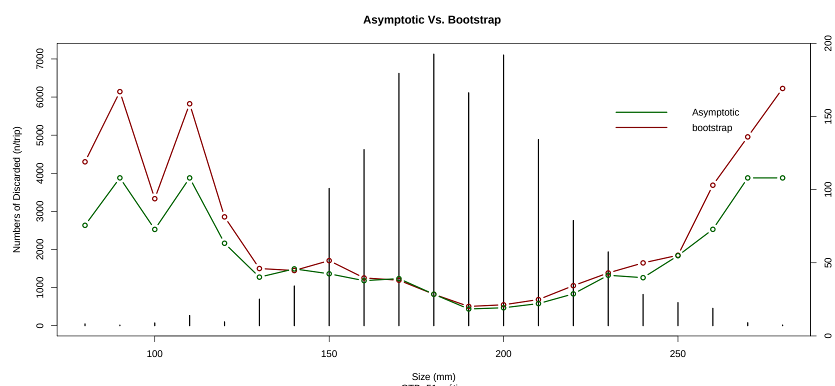

Figure 1: Asymptotic vs. bootstrap estimations for the original sampling scenario in OTB-51. CV values from the boostrap are clearly higher than the asymptotic. Difference are positively related with abundance.

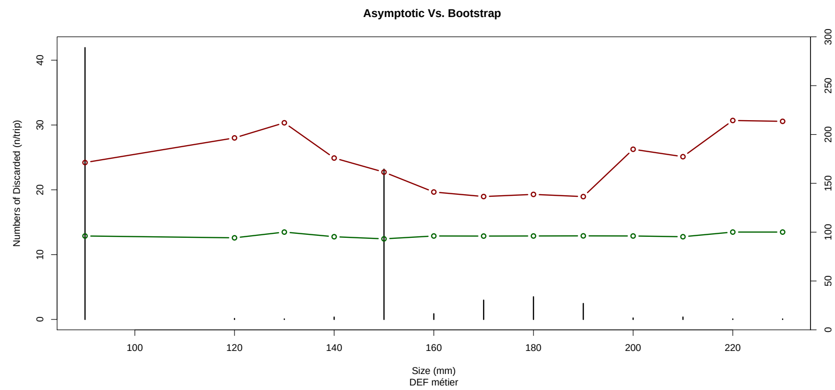

Figure 2: Asymptotic vs. bootstrap estimations for the original sampling scenario in `DEF`-Southern. CV values from the boostrap are clearly higher than the asymptotic. Difference are not clearly related with abundance in this case.

3.3 Study on the sources of variability

Figure **??** summarize the study of the effect of each source of error (trip, haul, intrahaul variability) on the uncertainty of discard estimations in `OTB-51`.sim_{ALL} is refered to the standar simulation where all identified sources of variability are resampled, sim_T fixes trip and resamples between and within hauls,sim_H fixes hauls and resamples between trips and within hauls,and sim_W resamples on trips and hauls. **??** show the change in error estimations by lengthclassess. sim_{HW} shows the highest differences compared to sim_{ALL}, while sim_H and sim_W , yield closer values. In our study Trip has accounted for 42% of the global variability in discards estimations, a superior effect compared with Haul $\sim$ 11% and length sampling $\sim$ 4%. Results must be taken with caution as the combined effects only explains $\sim$ 58%, that means a strong dependence between the three processes.

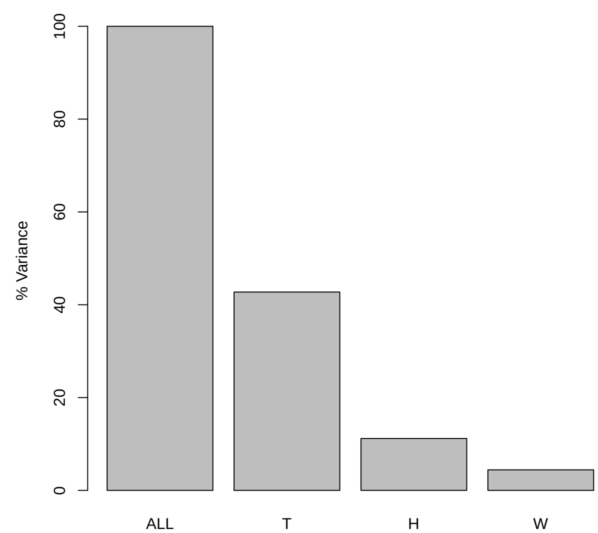

Figure 3: Relative variance from each source of variability. $ALL = sim_{ALL}$, $T = sim_T$,$H = sim_H$ and $W = sim_W$.

3.4 Simulated sampling scenarios

Figure **??** and **??** show the variation of ECVB values at different length sizes and sampling efforts for `OTB-51` and `DEF`-Southern respectively. The U-shape found in all `OTB-51` panels indicates higher ECVB in extreme sizes, due to the low presence of these individuals in the original dataset (see Figure 1). The reduction of ECVB values in this extreme sizes is consistent with the increment of the sampled trips,altought no clear improvement is found for central sizes. In the same way, extreme sizes are more sensitive to the change in sampling effort within trip, altougth the effect of this simulations is smaller than the change in number of trips.Figure **??** summarize the change in the estimated error for every simulations pooling all length sizes.The change of results is much more visible in the first simulations, while the effect of changing sampling effort within trip less pronounced.

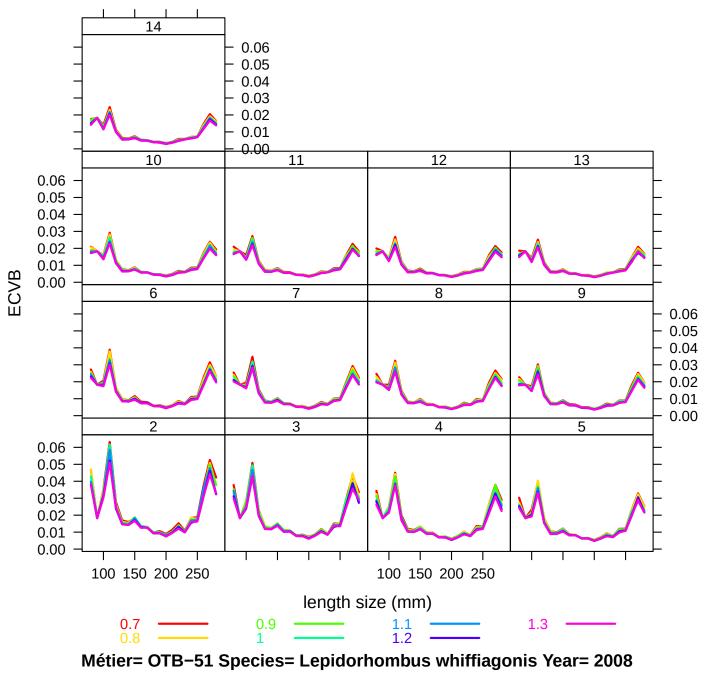

Figure 4: ECVB values obtained for different `OTB-51` sampling scenarios at length size level.

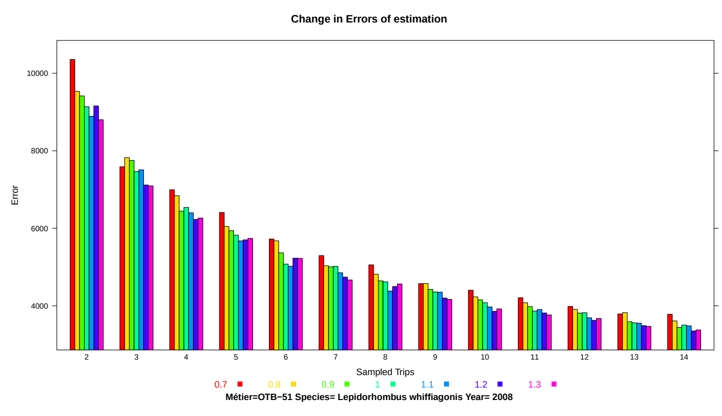

Figure 5: Change in error estimation for each of the simulated sampling scenarios for `OTB-51`.

3.5 Outlier detection

3.5.1 Approach 1: Linear approximation for outlier assessment

Table **??** and Figure **??** show the results of the Linear approximation for detecting outliers in the mediterranean data. Highest values in the estimates are

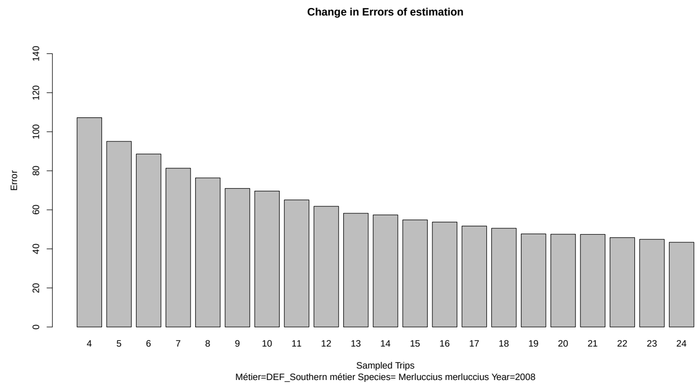

Figure 6: Change in error estimation for each of the simulated sampling scenarios for DEF-Southern.

found for Coefficients β_1 and β_2, linked with two of the three trips presenting discards. Low and similar coefficent values are found from β_3 to β_{14}, inluding the coefficient β_5 which is linked to the third trip presenting discards, altought in less amounts than trip 1 and trip 2.

	β_i	Std. Error	t value	Pr(>|t|)
Trip.1	86.6117	1.4487	59.79	0.0000
Trip.2	37.5391	1.4662	25.60	0.0000
Trip.3	1.9588	1.4567	1.34	0.1788
Trip.4	6.1049	1.4810	4.12	0.0000
Trip.5	3.5282	1.4794	2.38	0.0171
Trip.6	5.0333	1.4720	3.42	0.0006
Trip.7	3.0154	1.4669	2.06	0.0399
Trip.8	3.2091	1.5355	2.09	0.0367
Trip.9	5.9862	1.4554	4.11	0.0000
Trip.10	6.5967	1.5010	4.39	0.0000
Trip.11	3.8831	1.4740	2.63	0.0085
Trip.12	5.3439	1.4465	3.69	0.0002
Trip.13	3.7652	1.4704	2.56	0.0105
Trip.14	3.4034	1.5298	2.22	0.0262

Table 4: Coefficients β_i from the linear aproximationa and Errors of estimation.

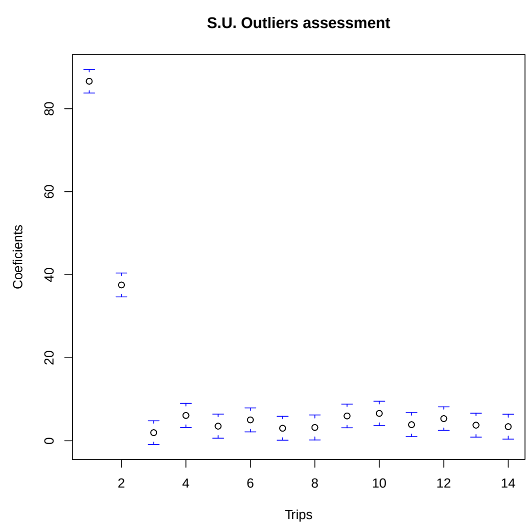

Figure 7: Coefficients from the linear aproximation.