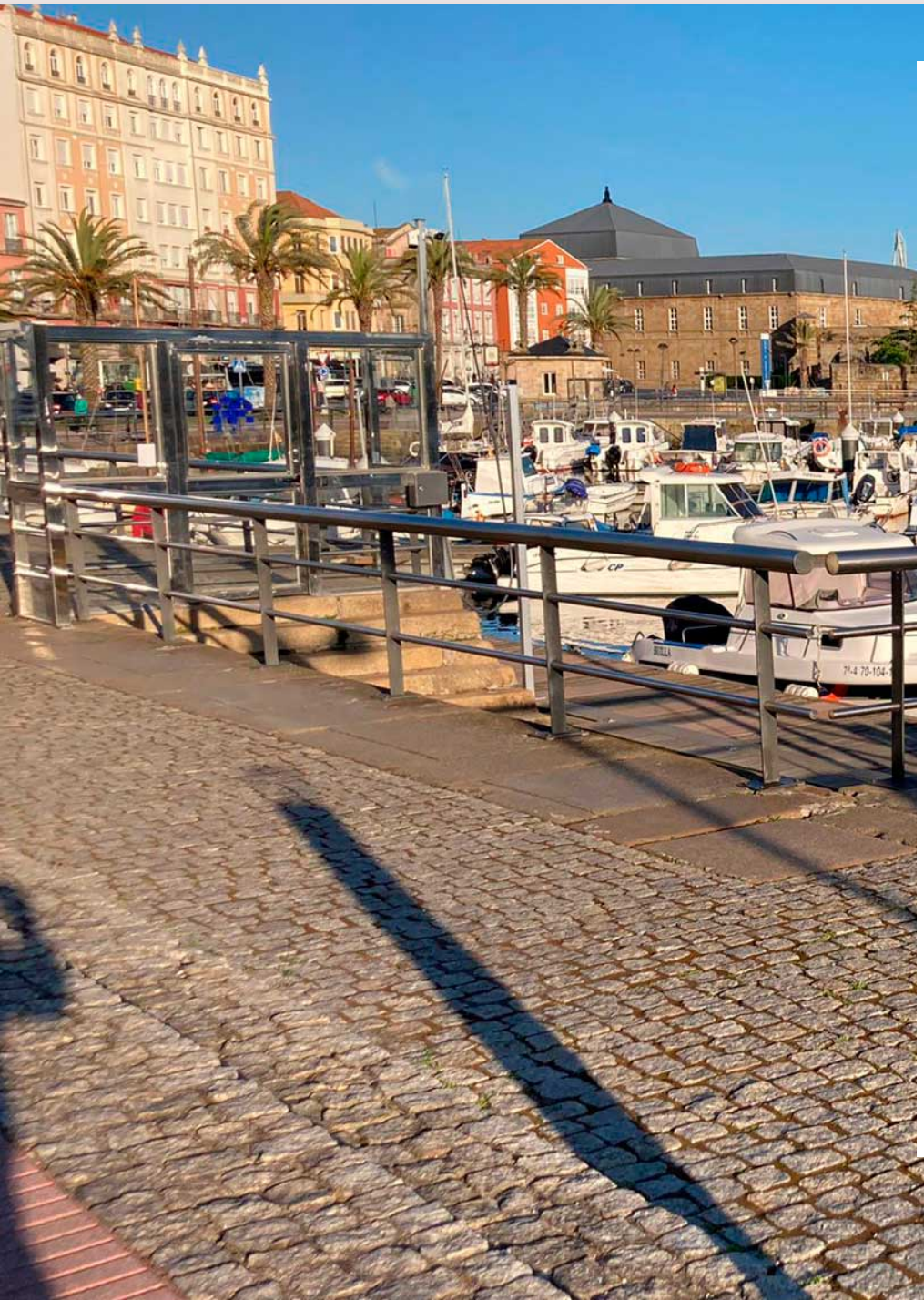


Anomaly detection using the qcr package

Salvador Naya, Javier Tarrío Saavedra, Miguel Flores, Rubén Fernández Casal

X Xornada de Usuarios de R en Galicia

Santiago de Compostela (Spain), 18/10/2023

UNIVERSIDADE DA CORUÑA citic

UNIVERSIDADE DA CORUÑA

ESCUELA POLITÉCNICA NACIONAL

ESCUELA POLITÉCNICA NACIONAL

Contents

1. `AnomalyDetection`: Loess.
2. `autoencoder`: Autoencoders.
3. `IsolationForest`: algoritm Isolation Forest.
4. `stats`: algoritm k-means.
5. `qcr` **(2021): Quality Control Review.**
   - **Flores, M., Fernández-Casal, R., Naya, S., & Tarrío-Saavedra, J. (2021). Statistical Quality Control with the qcr Package. R Journal, 13(1).**
   - **Miguel Flores, Fernandez-Casal R, Naya S, Tarrio-Saavedra J (2022). qcr: Quality Control Review. R package version 1.4. https://CRAN.R-project.org/package=qcr**

qcr: Quality Control Review

Univariate and multivariate SQC tools that completes and increases the SQC techniques available in R. Apart from integrating different R packages devoted to SQC ('qcc','MSQC'), provides nonparametric tools that are highly useful when Gaussian assumption is not met. This package computes standard univariate control charts for individual measurements, 'X-bar', 'S', 'R', 'p', 'np', 'c', 'u', 'EWMA' and 'CUSUM'. In addition, it includes functions to perform multivariate control charts such as 'Hotelling T2', 'MEWMA' and 'MCUSUM'. As representative feature, multivariate nonparametric alternatives based on data depth are implemented in this package: 'r', 'Q' and 'S' control charts. In addition, Phase I and II control charts for functional data are included. This package also allows the estimation of the most complete set of capability indices from first to fourth generation, covering the nonparametric alternatives, and performing the corresponding capability analysis graphical outputs, including the process capability plots.
See Flores et al. (2021) <doi:10.32614/RJ-2021-034>.

Version:	1.4
Depends:	R (≥ 2.10), qcc, fda.usc, mvtnorm, MASS
Suggests:	rmarkdown, knitr
Published:	2022-03-02
Author:	Miguel Flores [aut, cre], Ruben Fernandez-Casal [aut], Salvador Naya [aut], Javier Tarrio-Saavedra [aut]
Maintainer:	Miguel Flores <ma.flores at outlook.com>
BugReports:	https://github.com/mflores72000/qcr/issues
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL:	https://github.com/mflores72000/qcr
NeedsCompilation:	no
Materials:	README NEWS
CRAN checks:	qcr results

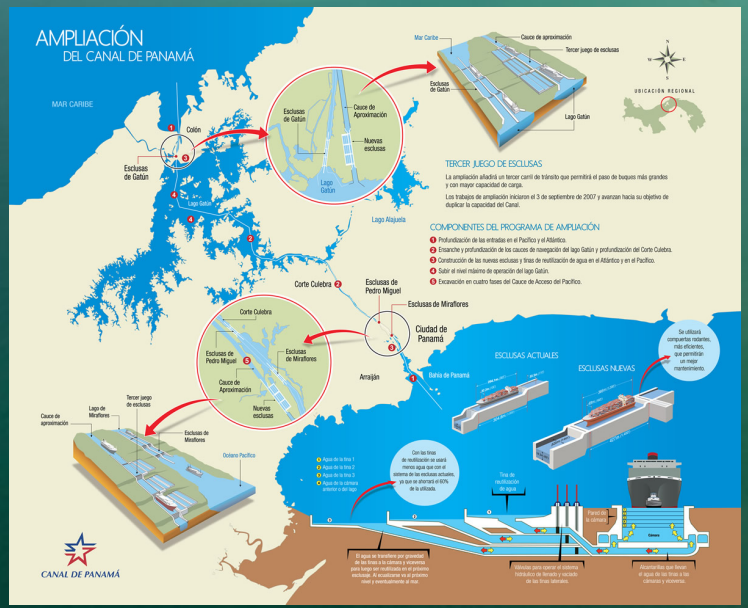
Documentation:

Reference manual: qcr.pdf

Downloads:

Package source:	qcr_1.4.tar.gz
Windows binaries:	r-devel: qcr_1.4.zip, r-release: qcr_1.4.zip, r-oldrel: qcr_1.4.zip
macOS binaries:	r-release (arm64): qcr_1.4.tgz, r-oldrel (arm64): qcr_1.4.tgz, r-release (x86_64): qcr_1.4.tgz, r-oldrel (x86_64): qcr_1.4.tgz
Old sources:	qcr archive

Process control, Anomaly detection, Analysis of the capability of processes to meet specifications.



EXPANDED PANAMA CANAL:

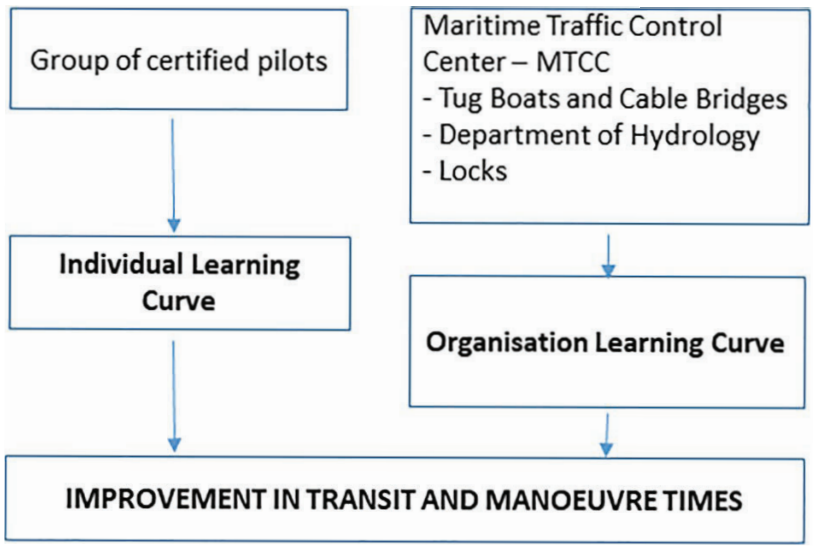
Locks: Cocolí (Pacific), Agua Clara (Atlantic).

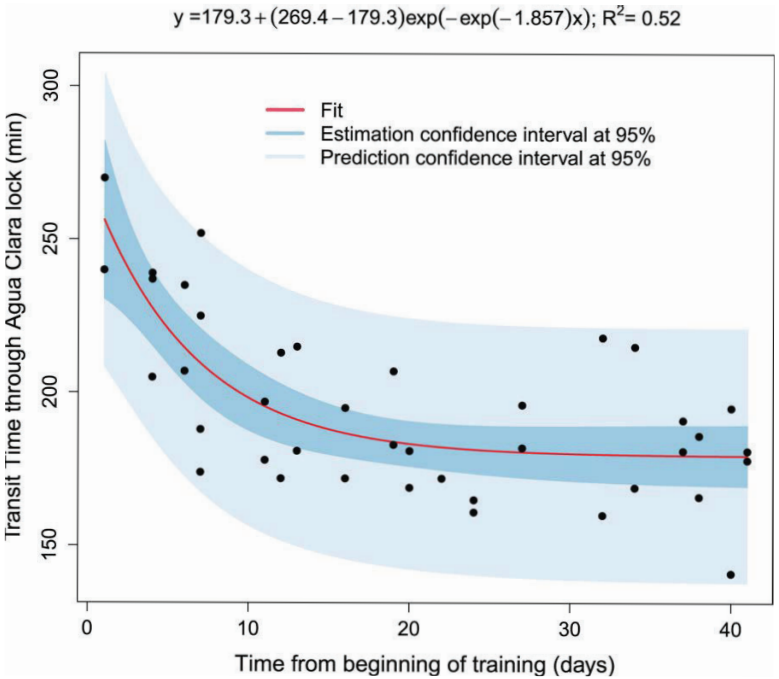
Vessels: LNG, LPG, Containers.

CTQ: time in transit along the locks (dir. N-S).

Identification of the learning pattern in the EPC

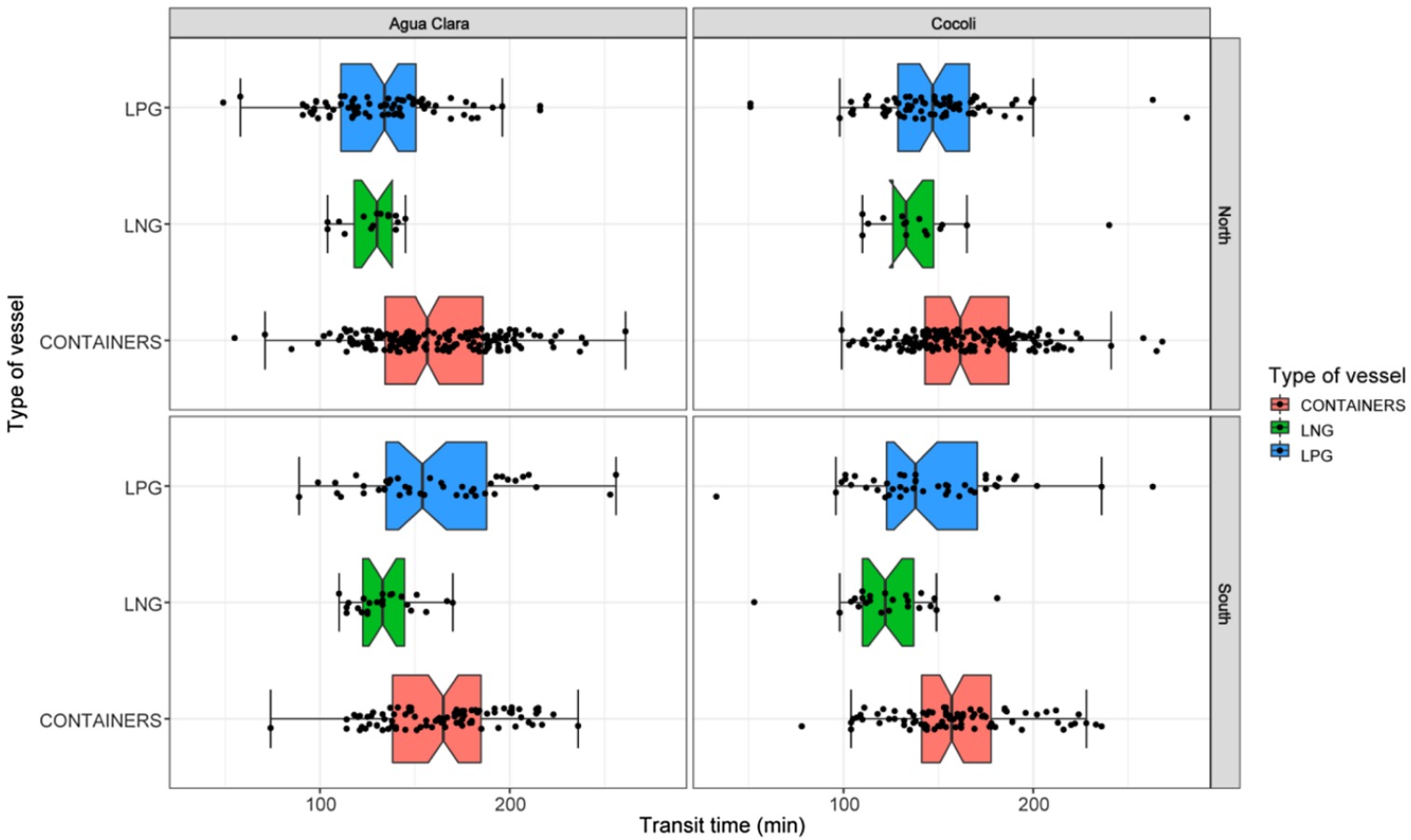
- Inauguration of the Expanded Panama Canal (EPC) in June 2016.
- **Goal: To identify and model a possible learning pattern in operators, pilots, organization.**
- **Critical variable for the quality of the EPC: Transit time of vessels through the two locks, Agua Clara (Atlantic) and Cocolí (Pacific).**





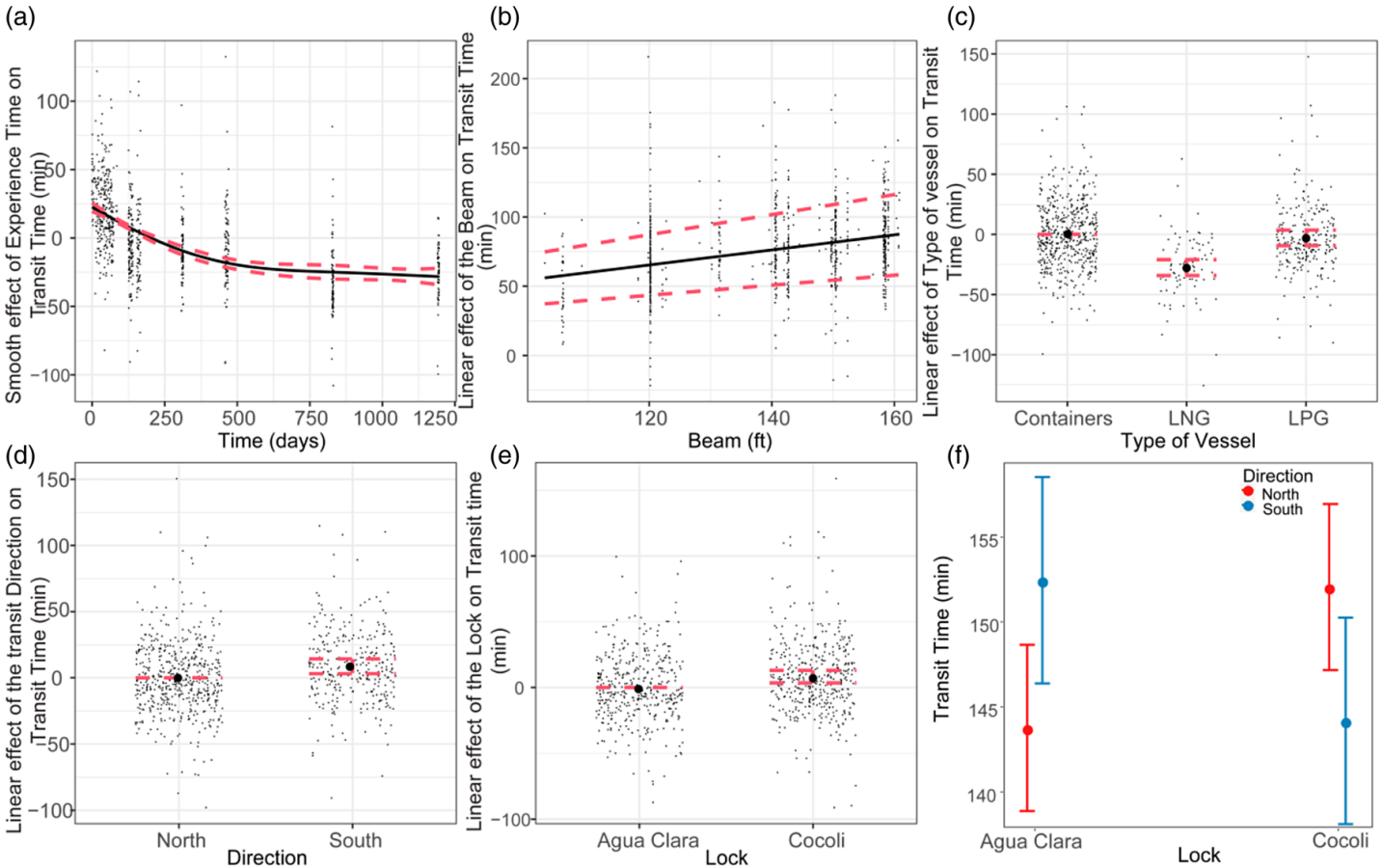
Identification of the learning pattern in the EPC

- Step 1: What variables affect the transit time through a lock?
- Descriptive analysis: Graphic ANOVA.
- Effect of vessel type, lock and direction.



Identification of the learning pattern in the EPC

- Second step: How is the effect of each variable?
- Semiparametric regression models: Generalized additive models (GAM).
- Multivariate regression models that allow the inclusion of linear and nonlinear (smooth) effects.



Univariate control charts

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_\alpha : \mu \neq \mu_0$$

$$\mathrm{UCL} = \mu_w + L\sigma_w$$

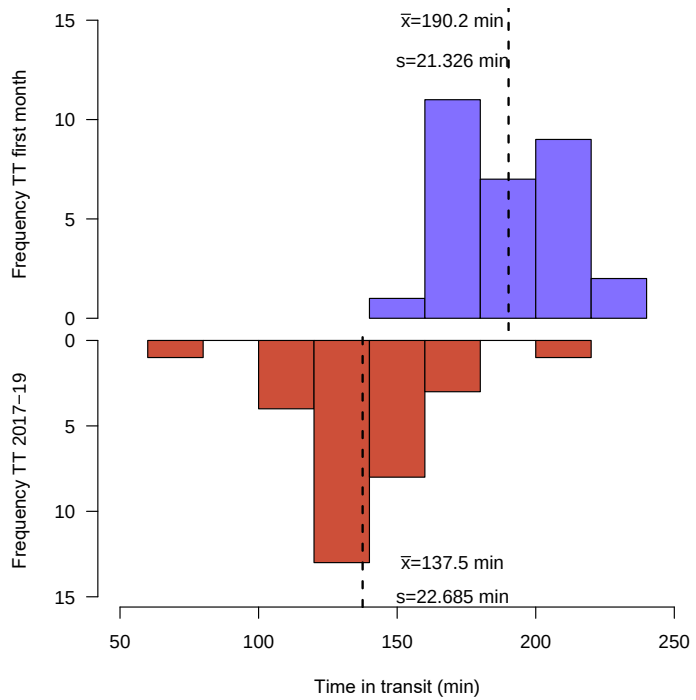
$$\mathrm{CL} = \mu_w$$

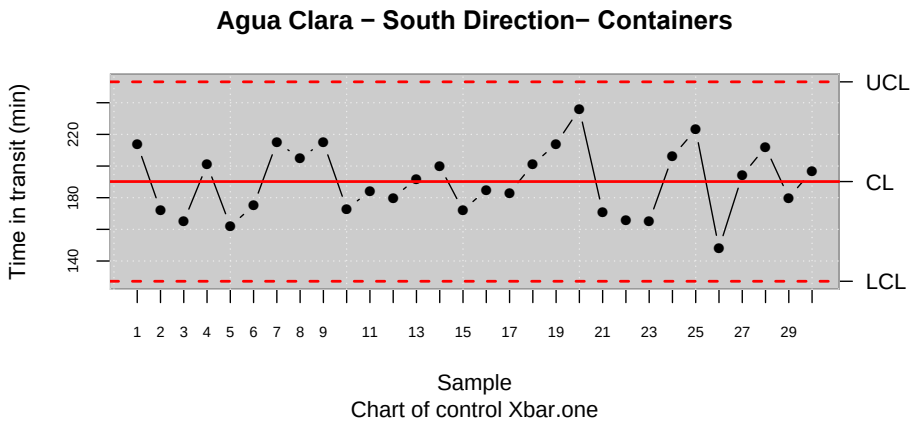
$$\mathrm{LCL} = \mu_w - L\sigma_w$$

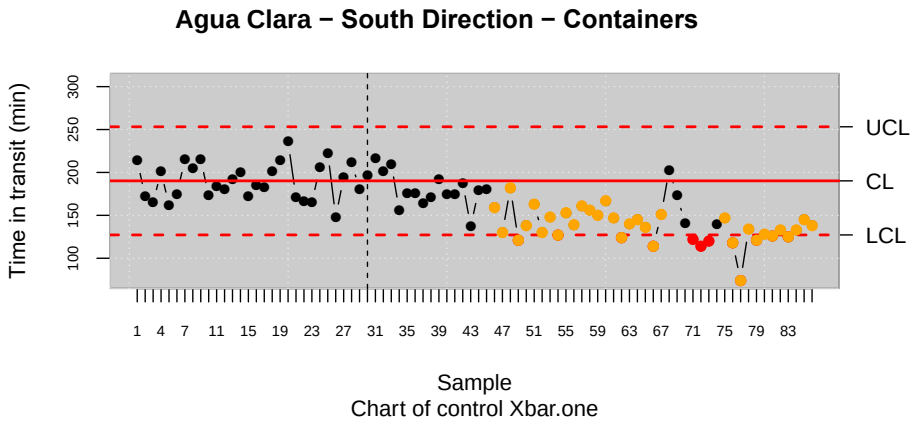
Univariate control: Individual measurement chart

- Transit of container vessels through Agua Clara, southbound.
- `qcd(data,var.index=1,sample.index=2,covar.index=NULL,covar.names=NULL,data.name=NULL, type.data=c("continuous","atribute","dependence"),sizes=NULL)`
- `qcs.one(x,center=NULL,std.dev=c("MR","SD"),k=2,conf.nsigma=3,limits=NULL,plot=FALSE,...)`

Process change identification: 16th observation.

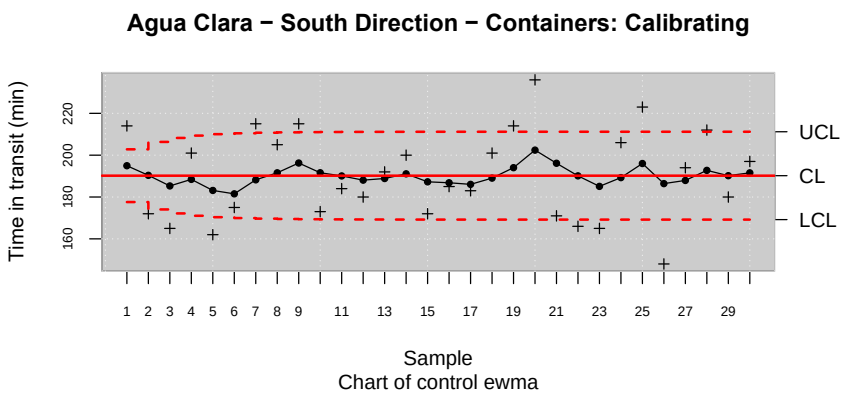


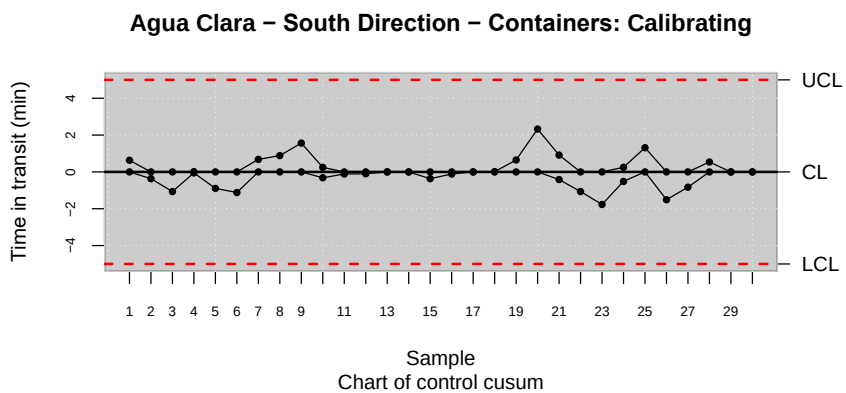


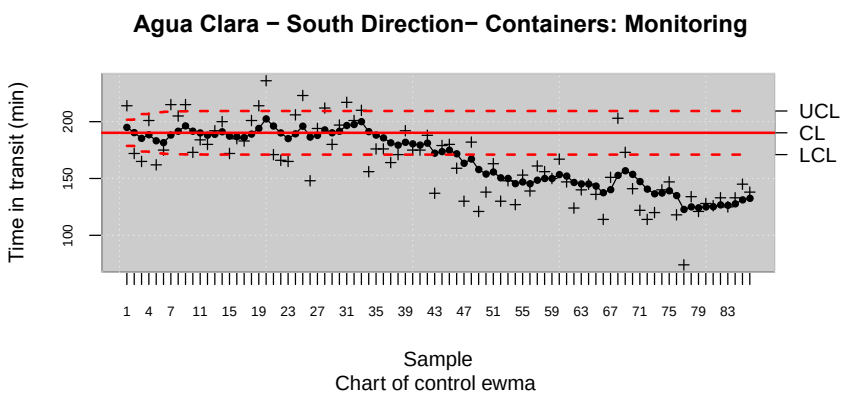


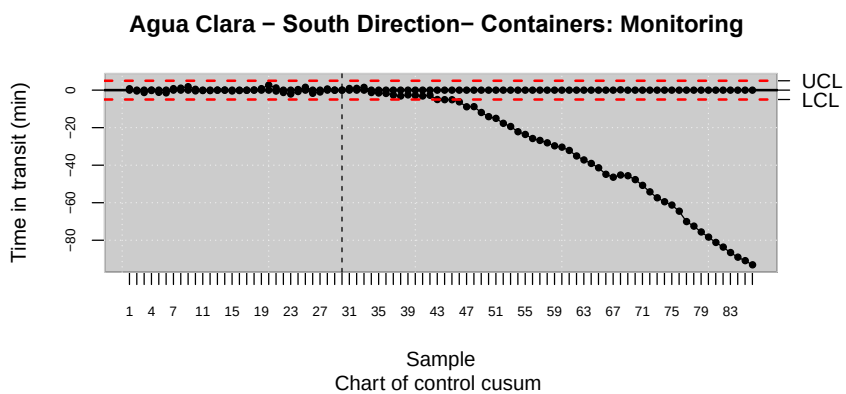
Univariate control: Memory control charts

- Transit of container vessels through Agua Clara, southbound.
- Detect changes of less than 2σ from mean.
- `qcs.ewma(x,center=NULL,std.dev=NULL,nsigma=3, lambda=0.2,plot=FALSE,...)`
- `qcs.cusum(x,center=NULL,std.dev=NULL,decision.interval = 5,se.shift = 1,plot = FALSE,...))`
- **Process change: 13th observation.**









Process Capability Analysis

- Are the Expanded Canal facilities capable of meeting the Panama Canal Authority's specifications?
- A process can be assumed capable when $PCR \geq 1{,}33$ (≈66 defects per million).
- $C_p\,(u, v) = \frac{d-u|\mu-m|}{3\sqrt{\sigma^2+v(\mu-T)^2}}$. The indices shown in the table are obtained by using this expression considering values of 0 or 1 for u and v: $C_p\,(0,0) = C_p$, $C_p\,(1,0) = C_{pk}$, $C_p\,(0,1) = C_{pm}$, $C_p\,(1,1) = C_{pmk}$, con $m = \frac{USL+LSL}{2}$ y $d = \frac{USL-LSL}{2}$.

Potential capability	$\hat{C}_p = \frac{USL-LSL}{6\hat{\sigma}}$
Actual capability with respect to the specification limits	$\hat{C}_{p,lower} = \frac{\hat{\mu}-LSL}{3\hat{\sigma}}$ $\hat{C}_{p,upper} = \frac{USL-\hat{\mu}}{3\hat{\sigma}}$ $\hat{C}_{pk} = \text{mín}\left[\frac{USL-\hat{\mu}}{3\hat{\sigma}}, \frac{\hat{\mu}-LSL}{3\hat{\sigma}}\right]$
Shifting of the mean with respect to the target	$\hat{C}_{pm} = \frac{\hat{C}_p}{\sqrt{1+\left(\frac{\hat{\mu}-T}{\hat{\sigma}}\right)^2}}$
C_{pk} correction for detecting deviations with respect to the target	$\hat{C}_{pkm} = \frac{\hat{C}_{pk}}{\sqrt{1+\left(\frac{\hat{\mu}-T}{\hat{\sigma}}\right)^2}}$

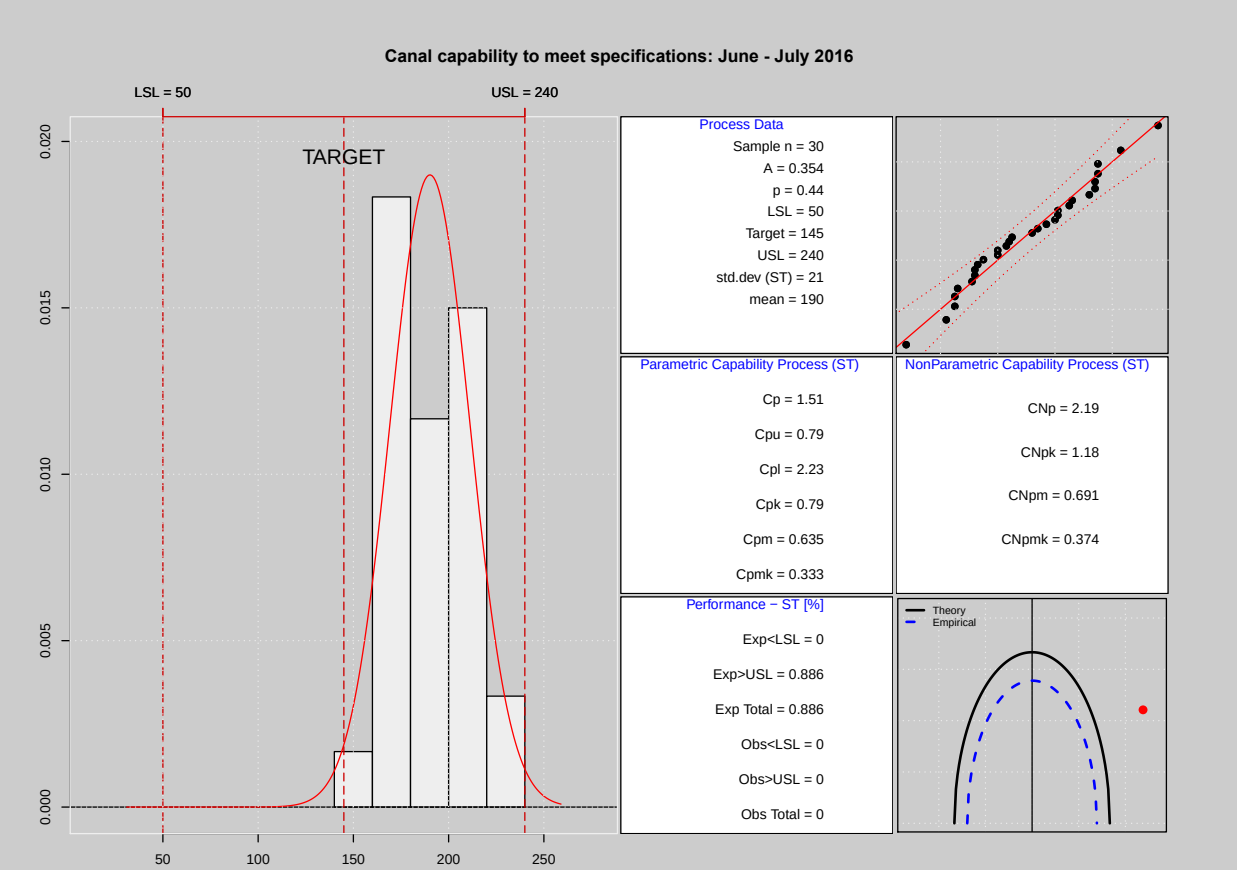
PCR from first to fourth generation, *USL*: upper specification limit, *LSL*: lower specification limit, μ: real mean, $\hat{\mu}$: estimated mean, y $\hat{\sigma}$: estimated standard deviation.

Capability graph - Nonparametric capability

- **Contour chart** of $C_p(u,v) = k$ as a function of δ (position) and γ (dispersion), with $\delta = \frac{\mu - T}{d}$ and $\gamma = \frac{\sigma}{d}$.
- **Capable process boundary:** $C_p(u,v)$ as a function of δ and γ according to $C_p(u,v) = \frac{1-u|\delta|}{3\sqrt{\gamma^2 + v(\delta)^2}}$.
- Thus, we solve the equation $C_p(u,v) = k$, plotting $\gamma = \sqrt{\frac{(1-u|\delta|)}{9k^2} - v\delta^2}$, $|\delta| \leq \frac{1}{u+3k\sqrt{v}}$, $(u,v) \neq (0,0)$. When $u = v = 0$ $(C_p = k)$, we have $\gamma = \frac{1}{3k}$ and $|\delta| \leq 1$.
- **Nonparametric version:** $\hat{C}_{Np}(u,v) = \frac{d - u|\hat{M} - m|}{3\sqrt{\left(\frac{U_p - L_p}{6}\right)^2 + v\left(\hat{M} - T\right)^2}}$, with U_p as estimates of $F_{99,865}$ y L_p de $F_{0,135}$, being $\hat{M}$ the median estimates, M.
- **Canal Authority Specifications:** $USL = 240min$, $LSL = 50min$.

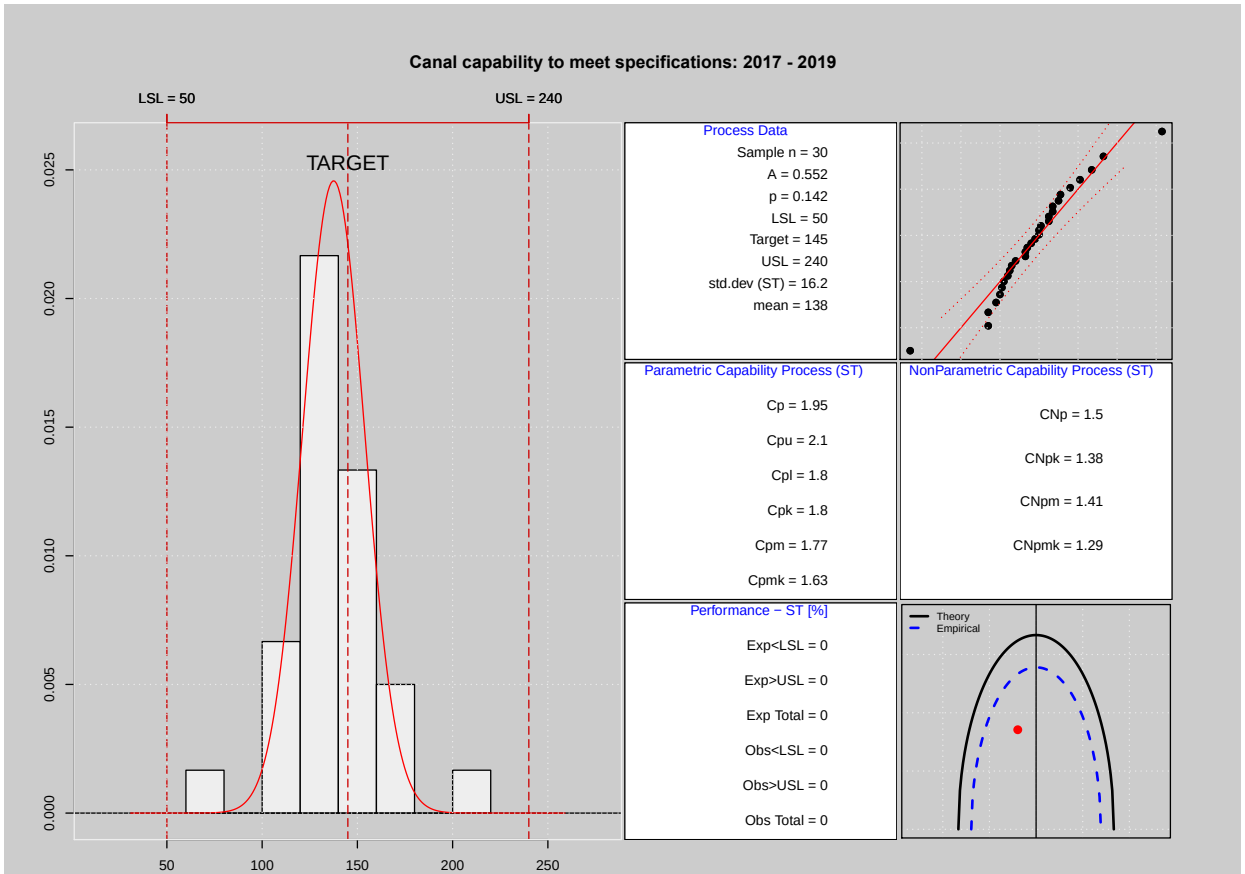
July 2016 transit time capability analysis

- $C_p > 1{,}33$, but C_{pk}, C_{pm}, $C_{pkm} < 1{,}33$. Potentially capable process but not in practice.
- Off-center process (see histogram, capability chart).
- `qcs.ca(object, limits=c(lsl=-3,usl=3), target=NULL, std.dev=NULL, nsigmas=3, confidence=0.9973, plot=TRUE, main=NULL,...)`



Transit time capability analysis 2017-2019

- C_p, C_{pk}, C_{pm}, $pkm > 1{,}33$. Process capable of meeting specifications.
- Centered process (see histogram, capability chart).
- **A learning pattern is identified for the facilities and pilots to meet management specifications.**



Multivariate process control: T^2

- $H_0 : \boldsymbol{\mu}_i = \boldsymbol{\mu}_0, \forall i$ vs $H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_0$, **system under control?** In order to answer this question, different statistics have been developed.
- **Hotelling T^2:**

$$T_i^2 = n\left(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_0\right)^\top \Sigma_0^{-1}\left(\bar{\mathbf{x}}_i - \boldsymbol{\mu}_0\right)$$

  where $\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \ldots, \bar{x}_{ip})$ is the vector of means corresponding to p features of the i-th subsample.
- If the process is under control $(\boldsymbol{\mu}_i = \boldsymbol{\mu}_0)$, α is the probability that T_i^2 exceeds the critical value $\mathcal{X}^2_{p,\alpha}$. If $T_i^2 > \mathcal{X}^2_{p,\alpha}$ there is a signal out of control.
- For individual measurements $UCL = \frac{p(m+1)(n-1)}{m^2 - mp} F_{\alpha,p,m-p}$ and $LCL = 0$.

Multivariate process control: MEWMA, CUSUM

- **MCUSUM chart** statistic:

$$G_i = max\left\{\left(G_{i-1} + a^\top (\mathbf{x}_i - \boldsymbol{\mu}_0) - 0{,}5D\right), 0\right\},$$

where $\boldsymbol{\mu}_0$ is the mean of the process under control, Σ_0 the variance-covariance matrix, $\boldsymbol{\mu}_1$ the mean of the process out of control, $D = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$ is a non-centrality parameter, and $a^\top = \frac{A}{D}$ with $A = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \Sigma_0^{-1}$.

- Alarm signal when $G_i > H$ (control limit).
- **MEWMA** statistic: From $Z_i = \Lambda \bar{\mathbf{x}}_i + (I - \Lambda) Z_{i-1}$,

$$T_i^2 = Z_i^\top \Sigma_{Z_i}^{-1} Z_i,$$

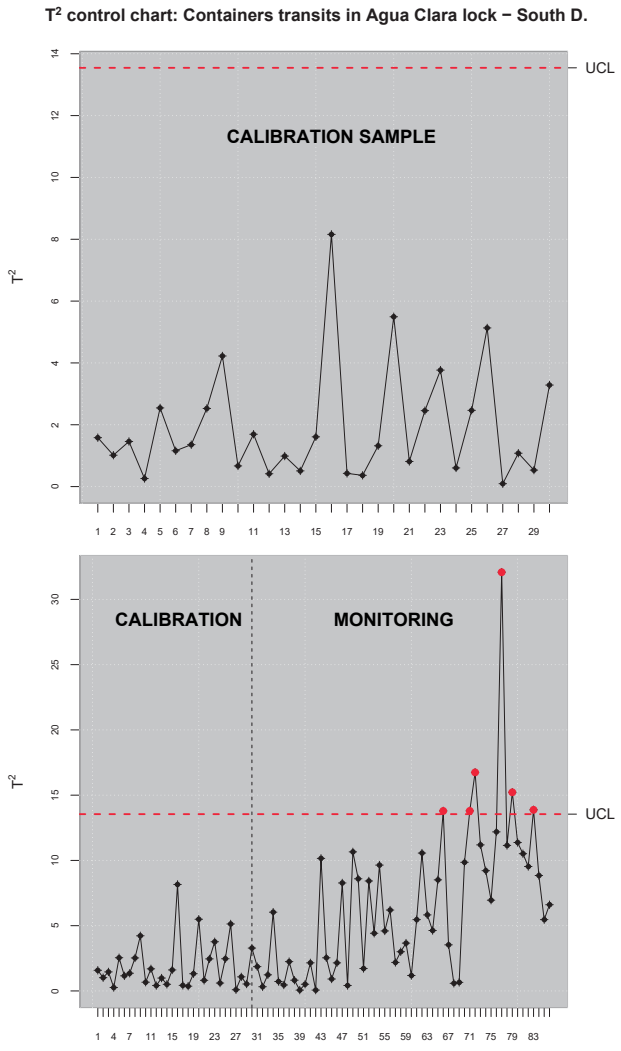
with $\Sigma_{Z_i}^{-1}$ being the inverse of the variance-covariance matrix of Z_i.

- Signal out of control when $T_i^2 > h$, selected taking into account the ARL.

Multivariate control charts: Hotelling T^2

- PHASE I (calibration): Estimation of control limits.
- PHASE II (monitoring): Does each new observation belong to the distribution of the calibration sample?
- There is a learning effect: The process has changed with respect to the calibration sample.
- Identification process out of control in the transit of the 36th vessel (2018).

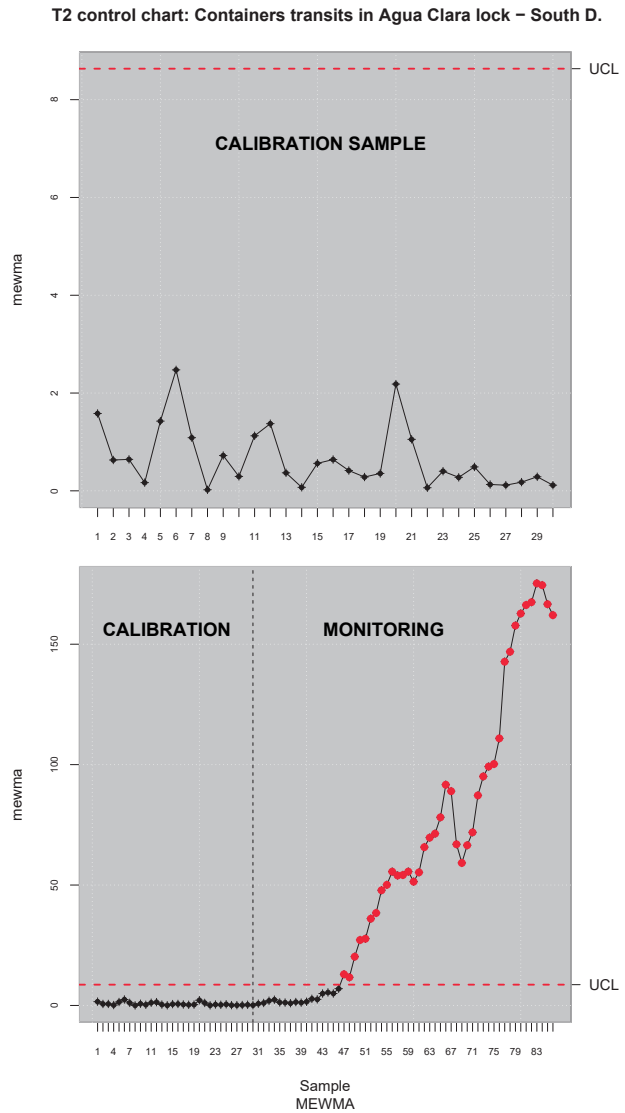
```
R> datos <-
as.matrix(Panama.sur.cont[1:86,7:8])
R> data.mqcd <- mqcd(datos)
R> res.mqcs.mot <- mqcs.t2(data.mqcd,Xmv=Xmv,
S=S, limits=res.mqcs$limits)
```



Multivariate control charts: MEWMA

- PHASE I (calibration): Estimation of control limits.
- PHASE II (monitored): Does each new observation belong to the calibration sample distribution?
- There is a learning effect: The process has changed with respect to the calibration sample.
- Identification process out of control in the transit of the 17th vessel (late 2016).

```
R> datos <-
as.matrix(Panama.sur.cont[1:86,7:8])
R> data.mqcd <- mqcd(datos)
R> res.mqcs.mot <-
mqcs.mewma(data.mqcd,Xmv=Xmv, S=S,
limits=res.mqcs$limits)
```



Case Study: Energy consumption control in stores located in Panama City

- Anomaly detection.
- Two CTQ variables: Energy consumption in air conditioning and lighting (kW).
- Daily consumption is measured.

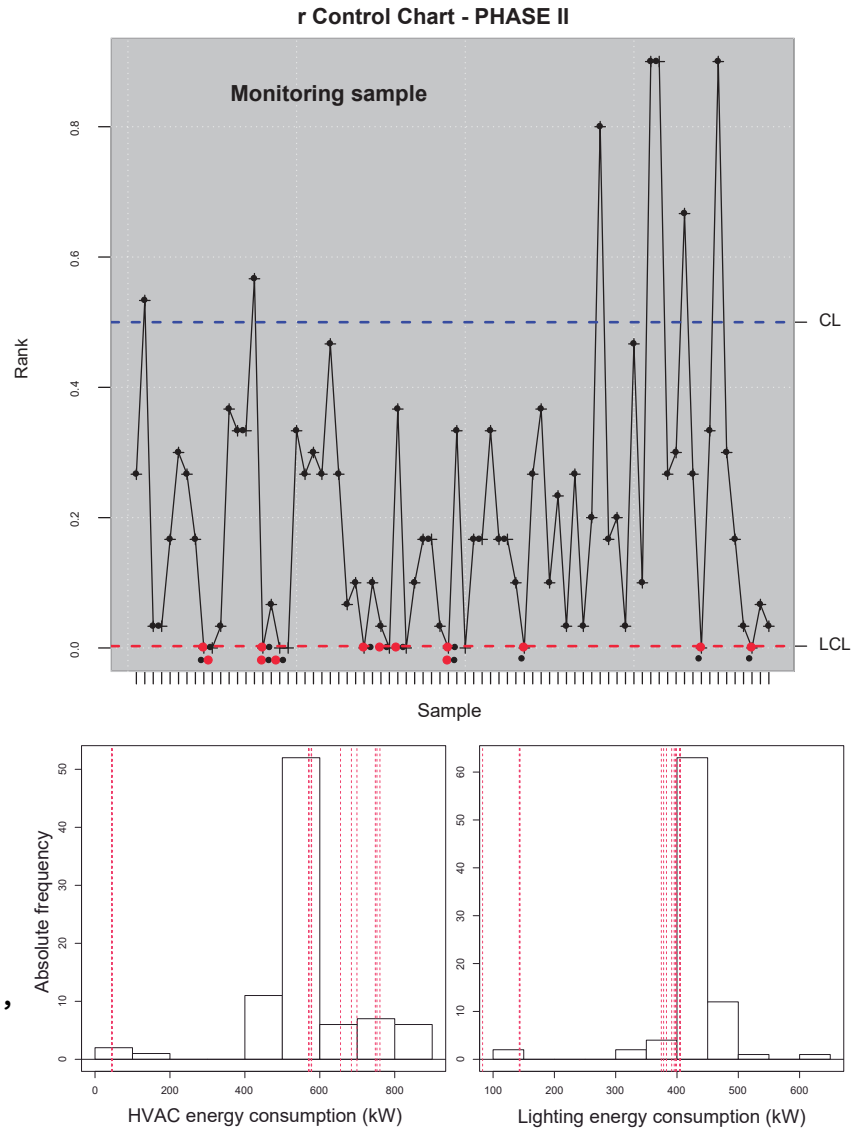
Nonparametric multivariate control charts

- **Based on data depth concept:**
  - Simplicial depth (Liu, 1990),
  - Mahalanobis depth (Mahalanobis, 1936),
  - Halfspace or Tukey depth (Tukey, 1975),
  - Likelihood depth (Fraiman, 1997), y
  - Random projection depth (Zuo, 2000).
- ***r* or Rank statistic**, alternative to individual measurement chart: $r_{G_m}(y) = \frac{\#\{D_{G_m}(Y_j) \leq D_{G_m}(y), j=1,\ldots,m\}}{m}$
- ***Q* statistic**, alternative to $\bar{x}$: $Q(G_m, F_n) = \frac{1}{n}\sum_{i=1}^{n} r_{G_m}(X_i)$
- ***S* statistic**, alternative to CUSUM: $S_n(G_m) = \sum_{i=1}^{n}\left(r_{G_m}(X_i) - \frac{1}{2}\right)$. Being $CL = 0$ and $LCL = -Z_\alpha\sqrt{n^2\frac{\left(\frac{1}{m}+\frac{1}{n}\right)}{12}}$.
- y: new multivariate observation; Y_j with $j = 1, \ldots, m$: calibration sample; $D_{G_m}(y)$: depth of y with respect to a calibration distribution, G_m; F_m: distribution of a sample to be monitored.

Rank chart application to energy consumption control

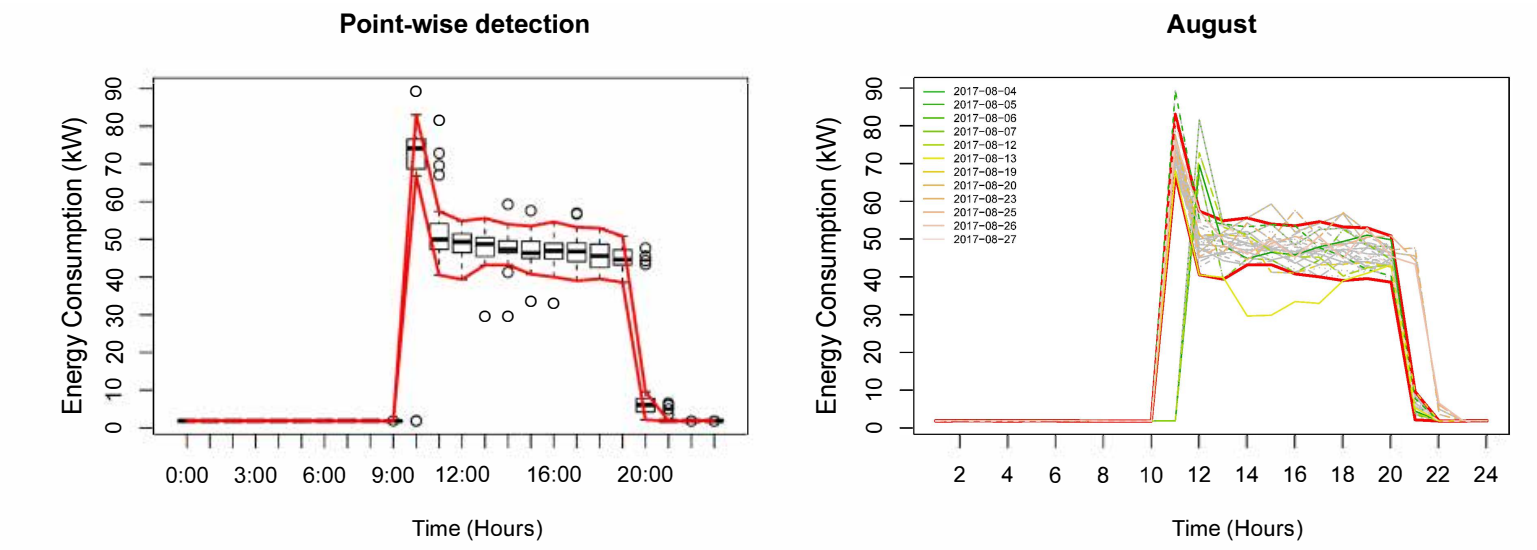
- PHASE II graph.
- Two variables: Daily HVAC and lighting consumption in a store in Panama →NOT GAUSSIAN, NOT AUTOCORRELATED.
- Saturday is open 1 hour longer than Monday-Friday and 2 hours longer than Sunday.
- Calibrated with Monday-Friday. Monitoring includes Saturdays and Sundays.
- Alarms: Sundays (low consumption), breakdowns and shutdowns, Saturdays (high consumption).

```
R> x<-as.matrix(Shop[c(44:dim(Shop)[1]),c(3,8)])
R> G<-as.matrix(Shop.week[c(1:30),c(3,8)])
R> data.npqcd<-npqcd(x, G)
R> res.npqcs<-npqcs.r(data.npqcd,method = "Tukey",
alpha = 0.0028)
R> plot(res.npqcs, title = r Control Chart")
```



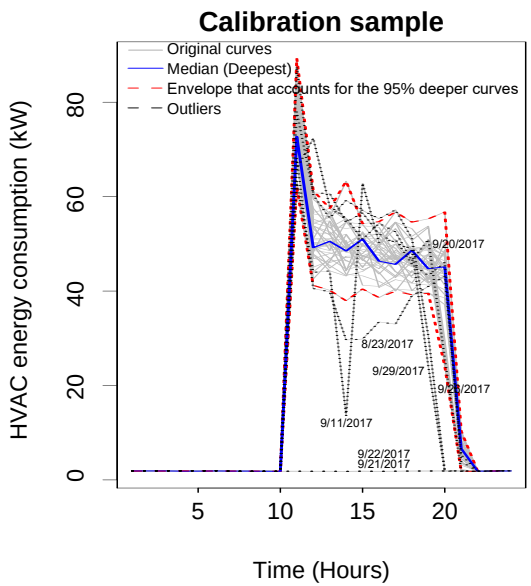
Control charts for Functional Data: HVAC case study

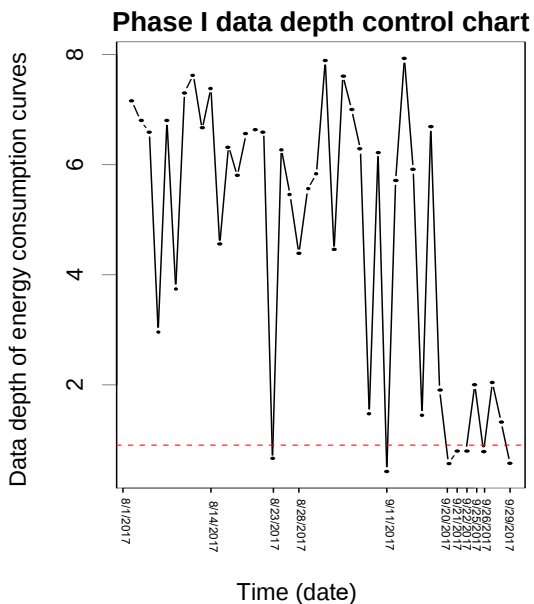
- September 11th: decrease in HVAC consumption towards the middle of the day. September 21st, 22nd and 30th: the shopping center was closed. September 27th: maintenance tests in the store facilities. September 29th: shutdown earlier than usual. From September 19th: the air conditioning is turned off half an hour earlier, i.e. there is a change of regulation in the HVAC system.
- Mid-October: a leak in the air conditioning circuit, energy consumption began to rise.
- November 1st: repair activities were carried out (↓ consumption and the initial consumption peak was avoided). Between November 17th and 20th, consumption increased again.



Control charts for Functional Data · Phase I

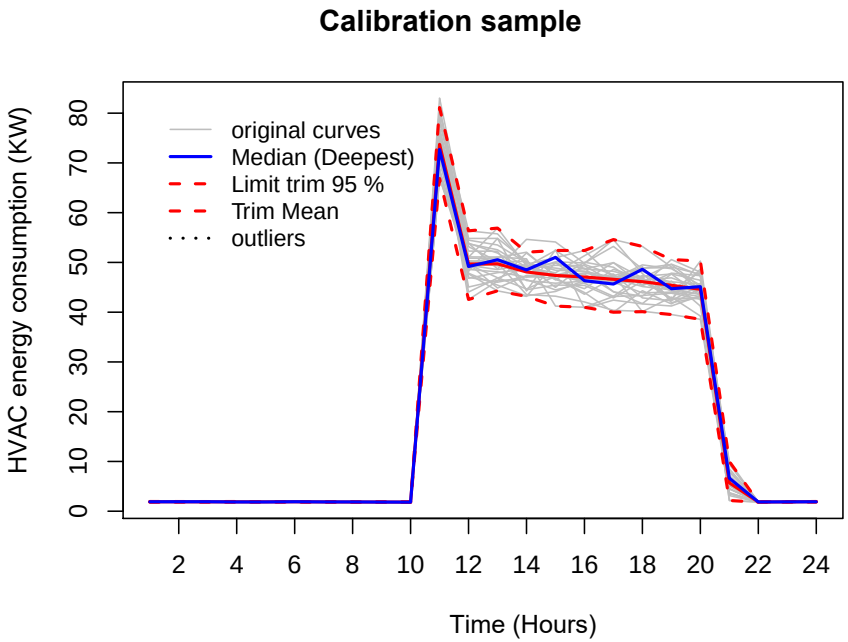
- PHASE I: A control chart based on data depth is proposed. With $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\}$, observations of a functional variable $\mathcal{X}$, the hypothesis $H_0 : \mathcal{X}_i(t) \overset{d}{=} \mathcal{X}_j(t), \forall i, j \in \{1, \ldots, n\}$, is tested against $H_a : \mathcal{X}_i(t) \overset{d}{\neq} \mathcal{X}_j(t)$, for some $i, j \in \{1, \ldots, n\}$
- The depth of each curve is calculated $D(\mathcal{X}_i)_{i=1}^n$ using Fraiman's depth, mode depth or random projections depth.
- The lower control limit (LCL) is estimated by a weighted bootstrap procedure (Febrero et al, 2008).

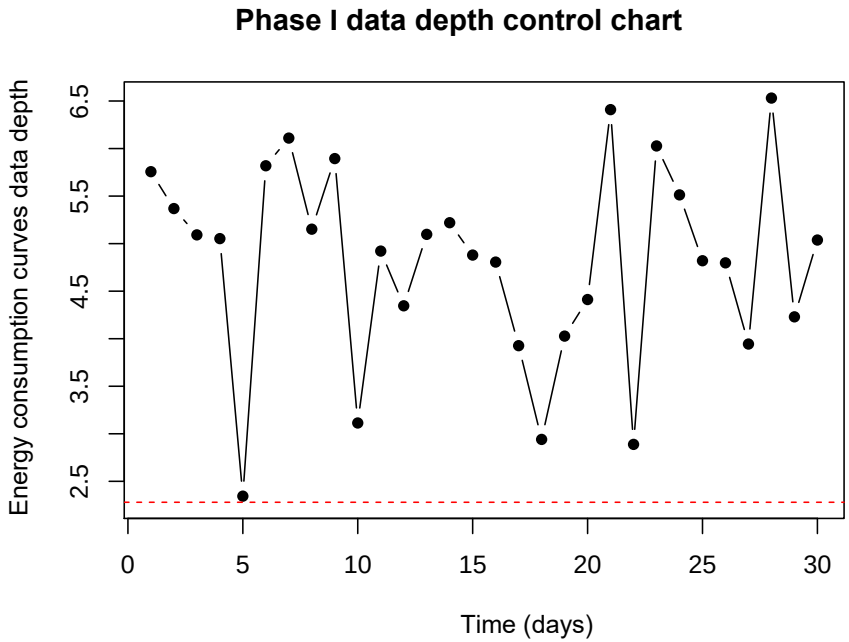




Control charts for Functional Data · Phase I

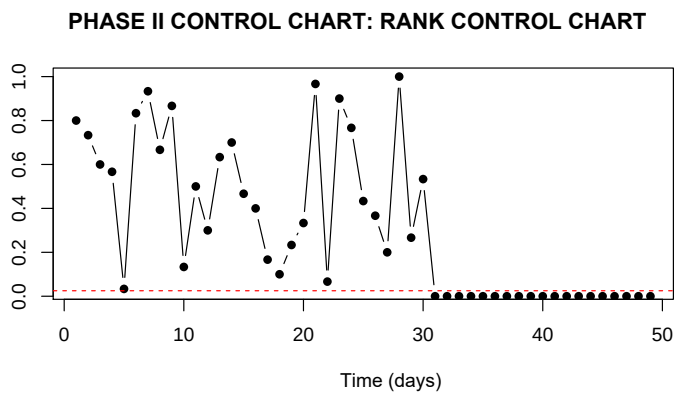
- The lower control limit (LCL) is estimated by a weighted bootstrap procedure (Febrero et al, 2008).
- An iterative process for outlier detection is applied.
- It is applied to the retrospective sample (August and September). All anomalies are detected. The representation of real curves helps to identify the assignable cause of each anomaly.

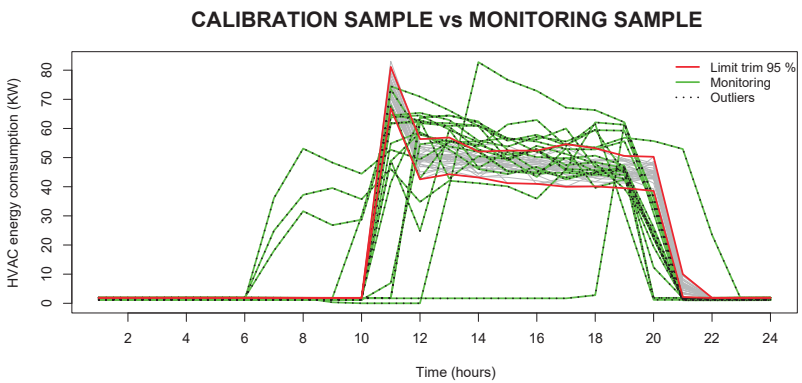




Control charts for functional data · Phase II

- PHASE II: Future observations are monitored by rank control charts from FDA depths.
- $\{\mathcal{X}_{n+1}(t), \mathcal{X}_{n+2}(t), \ldots, \mathcal{X}_m(t)\}$, with G distribution, are monitored, assuming that the calibration sample $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\}$ is distributed according to F. Thus, $H_0: F = G$ vs $H_1: F \neq G$ is tested.
- From $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\}$, the depths for calibration sample $D(\mathcal{X}_i)_{i=1}^n$, and for monitored sample, $D(\mathcal{X}_j)_{j=n+1}^m$, are obtained.
- The rank statistic $r_G(\mathcal{X}_{n+1}), \ldots, r_G(\mathcal{X}_m)$ is calculated, with respect to the calibration sample $\{\mathcal{X}_1(t), \mathcal{X}_2(t), \ldots, \mathcal{X}_n(t)\}$: $r_G(\mathcal{X}) = \frac{\#\{\mathcal{X}_i | D(\mathcal{X}_i) \leq D(\mathcal{X}), i=1,\ldots,n\}}{n}$
- Control chart: the central line is CL=0.5 and LCL= α. The process is monitored. IF $r_G(\mathcal{X}_j) \leq LCI$ for j, the process is out of control (anomaly). The original curves are shown with the envelope corresponding to the 99 % of the deepest (calibration).





“

Future functionalities to implement in qcr package

Multivariate approach: Anomaly detection using machine learning and bootstrap → Local Correlation Integral (LOCI) method (Energy Reports, 2023).

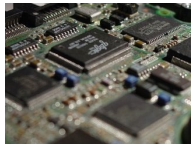
FDA approach: Anomaly detection using a FDA approach of LOCI method (Mathematics, 2023).

Bootstrap–LOCI anomaly detection method (Energy Reports, 2023)

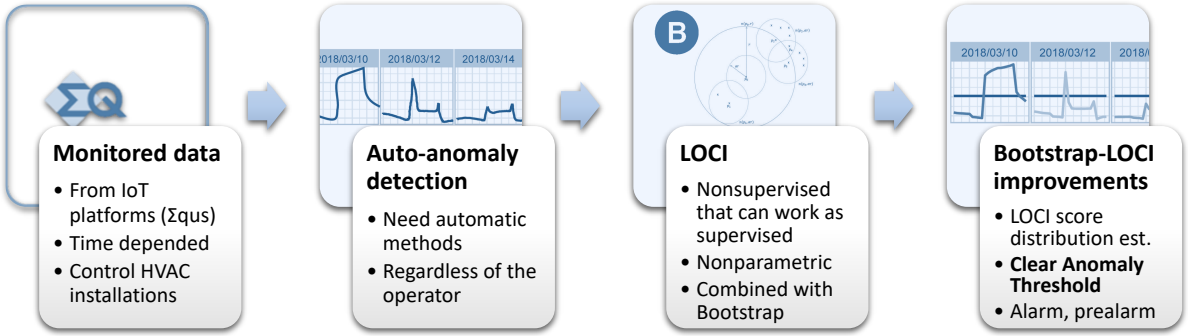
Bootstrap-LOCI data mining methodology for anomaly detection in buildings energy efficiency

The use of HVAC systems has become essential for work in industries.

Automatic identification of anomalies using Industrial 4.0 tools is crucial for energy efficiency in industrial plants.



With methodologies such as bootstrap-LOCI, automatic control of HVAC installations in industry is possible



Potential benefits

- Energy savings.
- Increased safety and quality of Heating Ventilation Air Conditioned (HVAC) service.
- Automation of anomaly detection processes in industry.

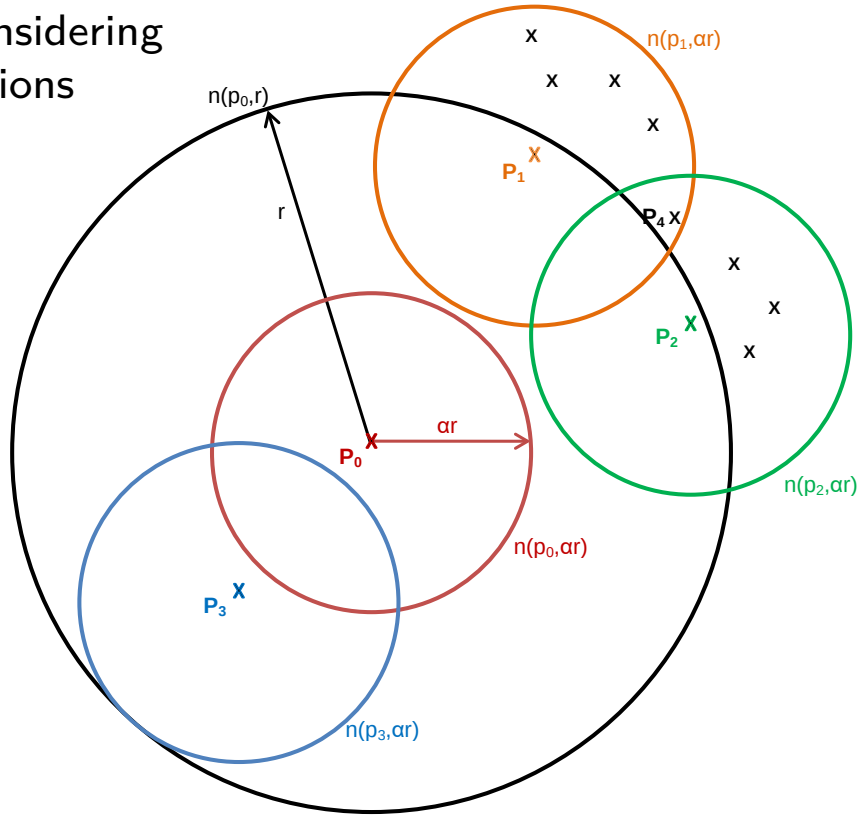
Innovations

1. Improvement of nonparametric LOCI cluster method to a supervised anomaly detection method.
2. Application of Bootstrap to estimate the LOCI score distribución $\Longrightarrow$ Estimation of a Threshold or Upper Control Limit to detect anomalies, alarms or pre-alarms.
3. Competitive method with respect to benchmark anomaly detection procedures (SVM, Log. Regr.).
4. Easy combination with visualization tools, implemented in R statistical software.

Bootstrap–LOCI anomaly detection method

From a sample $\mathbb{P}$ of size n, radii $r \in \mathbb{R}$ and considering $d(,)$ the Euclidean distance, for each observations $p_i \in \mathbb{P}$:

- $\mathcal{N}(p_i, r) = \{p \in \mathbb{P} : d(p, p_i) \leq r\}$, neighbourhood of p_i with r radius.
- $n(p_i, r) = |\mathcal{N}(p_i, r)|$, number of neighbours of p_i.
- $\alpha \in (0, 1]$, usaully 0.5.
- $\mathcal{N}(p, \alpha r)$ with $p \in N(p_i, r)$, number of elements in the sub-neighbourhood with αr radius.
- $MDEF(p_i, r, \alpha) = 1 - \frac{n(p_i, \alpha r)}{mean(n(p_i, r, \alpha))}$, the multi-granular deviation.
- $\sigma_{mdef}(p_i, r, \alpha) = \frac{\sigma(p_i, r, \alpha)}{mean(n(p_i, r, \alpha))}$, normalized standard deviation.



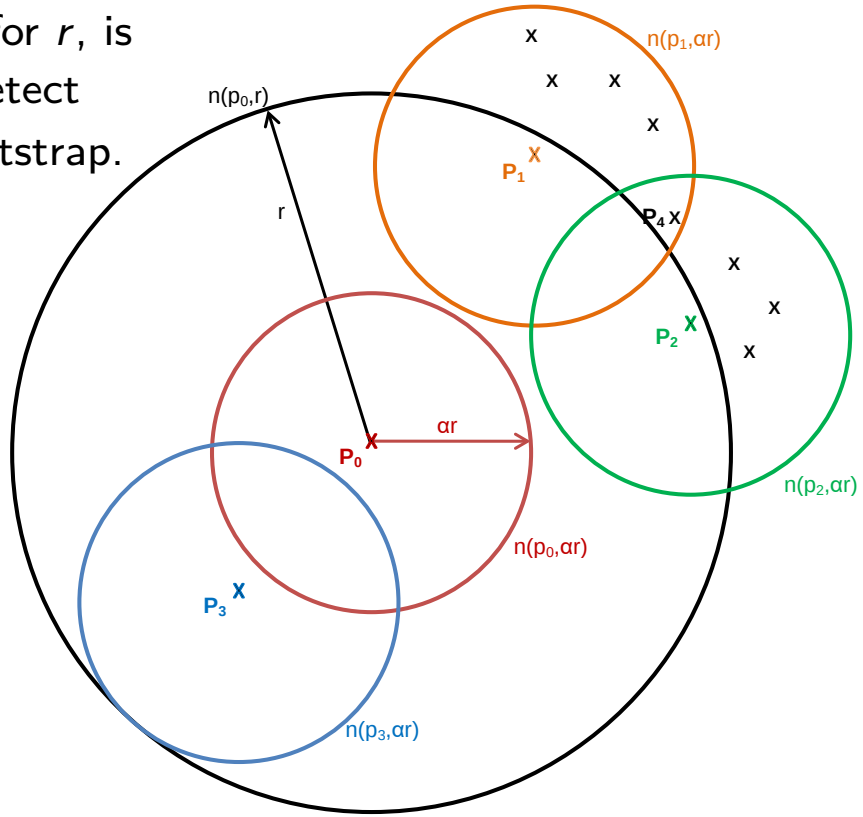
The observations $p_i \in \mathbb{P}$ are anomalies for any radii $r \in [r_{min}, r_{max}]$ if $MDEF(p_i, r, \alpha) > k(\sigma_{mdef}(p_i, r, \alpha))$, with $k > 0$.

Bootstrap–LOCI anomaly detection method

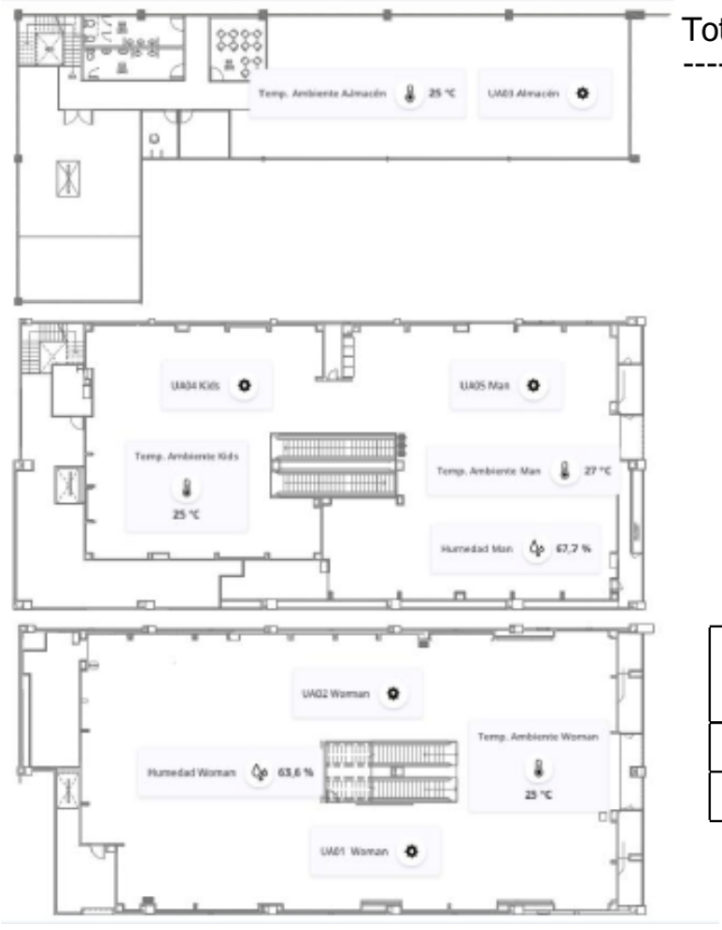
The score associated with the LOCI method, for r, is $\delta_{LOCI,r} = \frac{MDEF(p_i,r,\alpha)}{\sigma_{mdef}(p_i,r,\alpha)}$. It is the criterion to detect anomalies. Its distribution is estimated by bootstrap.

Algorithm 2: Bootstrap-LOCI

Result: Labelled observations
Select $\alpha \in (0,1]$;
Select $r \in [r_{min}, r_{max}]$;
For each $i = 1, ..., n$ throw X_i^* from F_n (using equation 1) ;
Obtain $X^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$;
Calculate $\delta^*_{LOCI,\alpha,r} = \delta_{LOCI,\alpha,r}(X^*, F_n)$ a from LOCI algorithm;
Repeat steps 3-5 B-times to obtain Bootstrap replicas.
$\delta^{*(1)}_{LOCI,\alpha,r}, ..., \delta^{*(B)}_{LOCI,\alpha,r}$;
Use these replicas to approximate the distribution of $\delta_{LOCI,\alpha,r}$.;
Obtain limit L from $\delta^{*(1)}_{LOCI,\alpha,r}, ..., \delta^{*(B)}_{LOCI,\alpha,r}$;
Apply LOCI algorithm to $X_1, X_2, \ldots, X_n$ and obtain $\delta_{LOCI,\alpha,r}$;
The decision if $\delta_{LOCI,\alpha,r} > L$ marks as an anomaly.



Case study: Energy efficiency and higrothermal comfort in a store of Panama



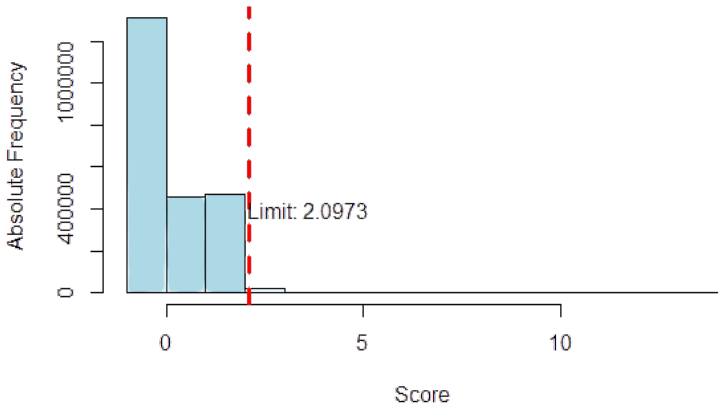
Total averages:

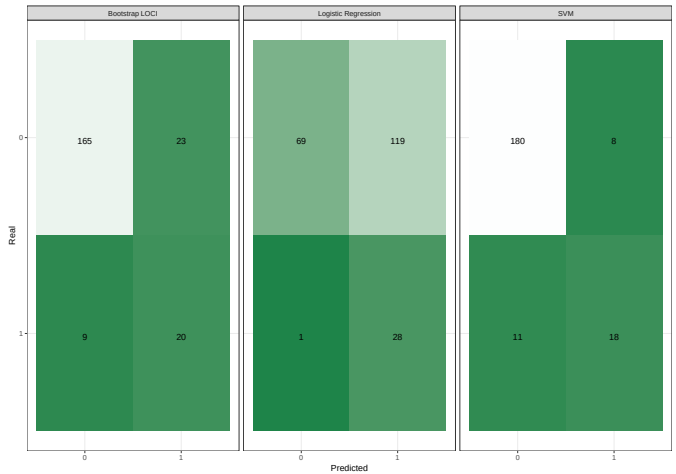
0.0267739208 -> Temperatura_(ºC)_Entrada_Agua_General
0.0266118808 -> Temperatura_Retorno_(ºC)_Agua_CL02_Ventas
0.0253941138 -> Temperatura_Impulsión_(ºC)_CL02_Ventas
0.0253431498 -> Temperatura_Ambiente_(ºC)_CL01_Ventas
0.0247264540 -> Temperatura_Retorno_(ºC)_Agua_CL01_Ventas
0.0236253646 -> Temperatura_Impulsión_(ºC)_CL01_Ventas
0.0232939155 -> Temperatura_Ambiente_(ºC)_CL02_Ventas
0.0154717839 -> Humedad_Relativa_(%)_Ventas
0.0132083763 -> Temperatura_Retorno_(ºC)_Agua_CL03_Almacén
0.0110942777 -> Temperatura_Ambiente_(ºC)_CL03_Almacén
0.0107792047 -> Temperatura_Impulsión_(ºC)_CL03_Almacén
0.0107327591 -> Analizador_Chiller_Potencia_Activa_(KW)
0.0102031855 -> Analizador_General_Potencia_Activa_(KW)

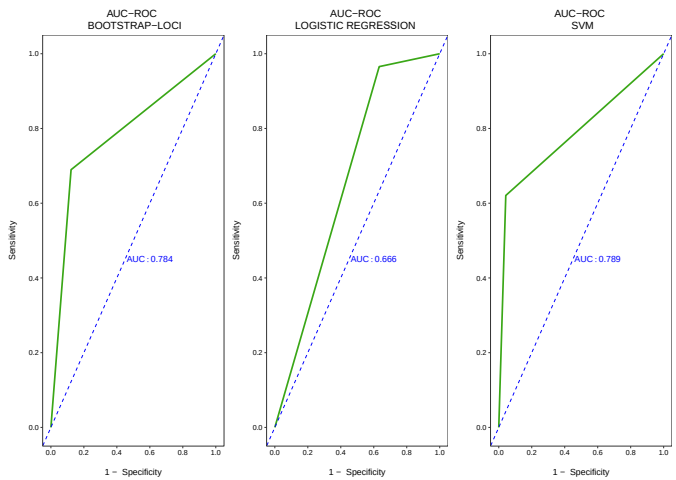
Set	Number of days	Number of anomalies	Total record
Training	217	28	5208
Testing	217	29	5208

Bootstrap–LOCI anomaly detection method

- LOCI parameters calculated from the training sample: 189 days that were considered non-atypical. $\alpha = 0{,}75$ is the smallest with which most anomalies are identified.
- We worked with 500 Bootstrap samples. Score distribution is estimated and the 99 percentile is fixed to detect anomalies.







- Bootstrap-LOCI has better performance than LOCI when score threshold is estimated empirically or from Chebyshev's Theorem.
- Bootstrap-LOCI has higher accuracy and sensitivity than Logistic Regression, detecting a significant less number of false alarms.
- Bootstrap-LOCI has a specificity higher than SVM (it classifies better the anomalous days), similar sensitivity, accuracy, and AUC.

Indicator	Bootstrap-LOCI Limit$= 2,0973$	Logistic regression stepwise	SVM polynomial kernel
TPR	0.8777	0.3670	0.9574
TNR	0.6897	0.9655	0.6207
ACC	0.8525	0.447	0.9124
BA	0.7837	0.6663	0.7891

“

References

Flores, M., Fernández-Casal, R., Naya, S., & Tarrío-Saavedra, J. (2021). Statistical Quality Control with the qcr Package. **R Journal**, 13(1).

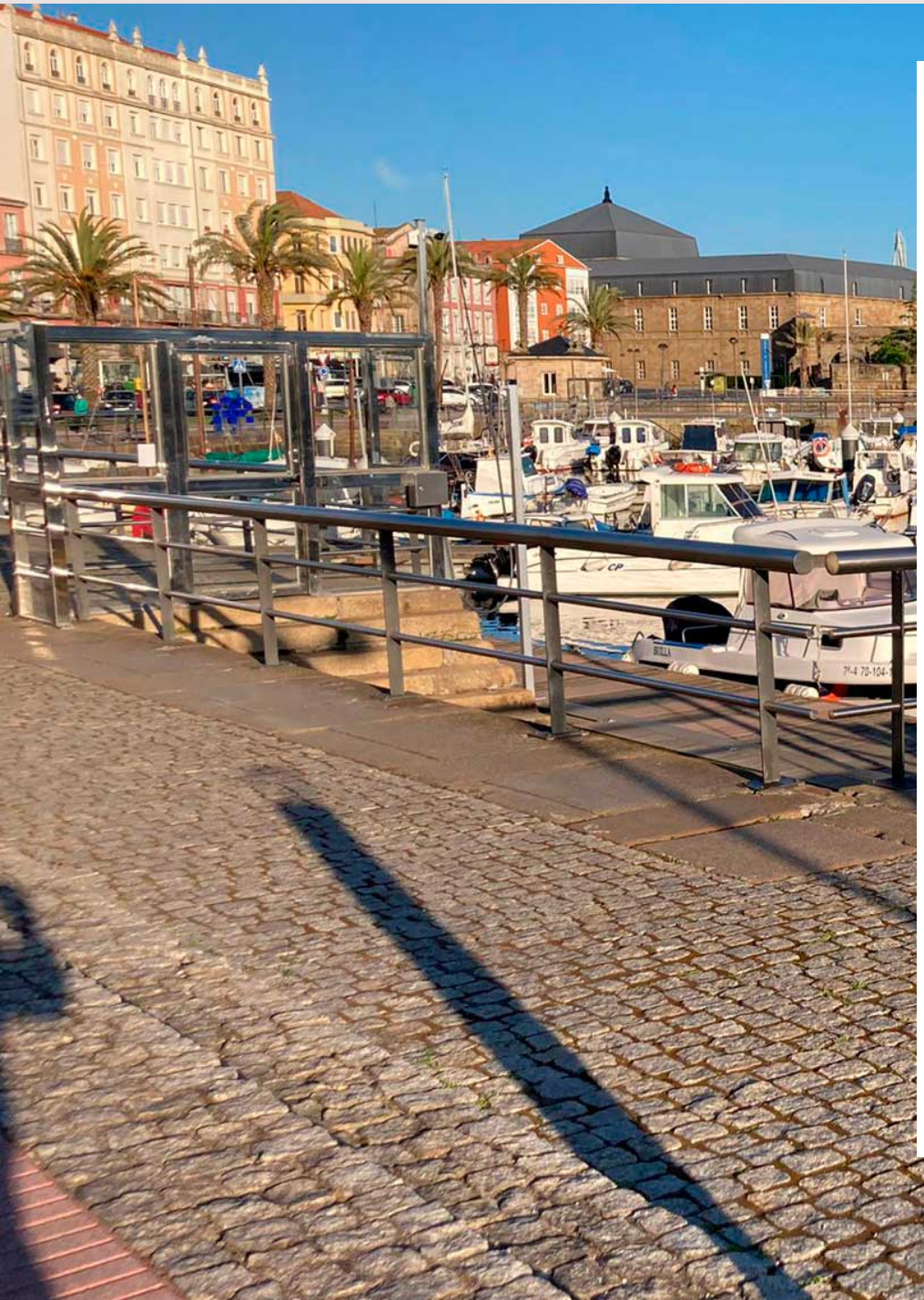
Flores, M., Naya, S., Fernández-Casal, R., Zaragoza, S., Raña, P., & Tarrío-Saavedra, J. (2020). Constructing a control chart using functional data. **Mathematics**, 8(1), 58.

Carral, L., Tarrío-Saavedra, J., Sáenz, A. V., Bogle, J., Alemán, G., & Naya, S. (2021). Modelling operative and routine learning curves in manoeuvres in locks and in transit in the expanded Panama Canal. **The Journal of Navigation**, 74(3), 633-655.

Flores, M., Moreno, G., Solórzano, C., Naya, S., & Tarrío-Saavedra, J. (2021). Robust bootstrapped Mandel'sh and k statistics for outlier detection in Interlaboratory Studies. **Chemometrics and Intelligent Laboratory Systems**, 104429.

Tobar, A., Flores, M., Castillo, S., Naya, S., Zaragoza, S., Tarrío-Saavedra, J. (2023). Bootstrap–LOCI data mining methodology for anomaly detection in buildings energy efficiency. **Energy Reports**, 10, 244-254. Sosa Donoso, J. R., Flores, M., Naya, S.,

Tarrío-Saavedra, J. (2023). Local Correlation Integral Approach for Anomaly Detection Using Functional Data. **Mathematics, 11(4), 815.**



Anomaly detection using the qcr package

Salvador Naya, Javier Tarrío Saavedra, Miguel Flores, Rubén Fernández Casal

X Xornada de Usuarios de R en Galicia

Santiago de Compostela (Spain), 18/10/2023

UNIVERSIDADE DA CORUÑA citic

UNIVERSIDADE DA CORUÑA

ESCUELA POLITÉCNICA NACIONAL

ESCUELA POLITÉCNICA NACIONAL