



An R package for batch process control and monitoring using Dual STATIS-Parallel Coordinates approach

Un paquete de R para control y monitoreo de procesos por lotes utilizando el enfoque Statis Dual-Coordenadas Paralelas

Ing. José Ascencio-Moreno¹

josdasce@espol.edu.ec

<https://orcid.org/0000-0002-6883-7195>

Blga. Miriam Vanessa Hinojosa-Ramos²

mvhinojo@espol.edu.ec

<https://orcid.org/0000-0002-4100-5284>

PhD. Francisco Vera Alcívar³

fvera@espol.edu.ec

<https://orcid.org/0000-0001-6541-7243>

PhD. Omar Ruiz-Barzola⁴

oruiz@espol.edu.ec

<https://orcid.org/0000-0001-8206-1744>

PhD. María Purificación Galindo-Villardón⁵

pgalindo@usal.es

<https://orcid.org/0000-0001-6977-7545>

MPC. Miriam Ramos-Barberán⁶

mvramosb@espol.edu.ec

<https://orcid.org/0000-0002-8915-6938>

Recibido: 1/9/2020, Aceptado: 1/11/2020

RESUMEN

El control estadístico multivariante de procesos para la producción por lotes generalmente toma en consideración características correlacionadas para la inspección del desempeño del proceso. En la literatura, los investigadores han utilizado varias técnicas estadísticas de forma individual para abordar esta inspección durante las fases de control y seguimiento. Nuevas estrategias han explorado la posibilidad de combinar dos técnicas con el fin de optimizar el control y el monitoreo del proceso por lotes, como el enfoque DS-PC. Este enfoque novedoso se refiere al uso de Statis Dual y Coordenadas Paralelas e implica una serie de varios pasos de protocolos y aplicaciones de fórmulas que son propensas a errores y consumen mucho tiempo. Utilizando la metodología que se encuentra en la literatura, el paquete DSPC para R

se desarrolló con el objetivo de ofrecer una herramienta simple para realizar el cómputo de Statis Dual rápidamente para las fases de control y seguimiento. Las

¹ Escuela Superior Politécnica del Litoral, Ecuador

² Escuela Superior Politécnica del Litoral, Ecuador

³ Escuela Superior Politécnica del Litoral, Ecuador

⁴ Escuela Superior Politécnica del Litoral, Ecuador

⁵ Universidad de Salamanca, España

⁶ Escuela Superior Politécnica del Litoral, Ecuador

salidas del paquete ofrecen visualizaciones gráficas para detectar comportamientos inusuales durante la producción a través de gráficos de control IS (Interestructura) y CO (Intraestructura). La salida también incluye el gráfico de coordenadas paralelas. Este paquete será útil para los profesionales interesados en la aplicación del enfoque DS-PC a cualquier industria de proceso por lotes a través de la automatización sugerida por defecto o la opción personalizada. Para familiarizar a los usuarios con esta estrategia, el paquete proporciona un conjunto de datos simulado de fabricación de bolsas de plástico industriales.

Palabras clave: producción por lotes, control, monitoreo, DS-PC, paquete

ABSTRACT

Multivariate statistical process control for batch production usually takes into consideration correlated characteristics for inspection of process performance. In literature, researchers have used several statistical techniques individually to address this inspection during the pilot and the monitoring phases. New strategies have explored the possibility of combining two techniques in order to optimize batch process control and monitoring, such as, DS-PC approach. This novel approach stands for Dual STATIS and Parallel Coordinates and involves a multi-step series of protocols and formula applications that are error-prone and time consuming. Using the methodology found in the literature, DSPC package for R was developed to deliver a simple tool to quickly compute Dual STATIS for pilot and monitoring phases. Outputs of the package offer graphic displays to detect unusual behavior during the production through IS (Interstructure) and CO (Intrastructure) control charts. Output also includes Parallel Coordinates plot. This package will be useful to practitioners interested in DS-PC approach application to any batch process industry through suggested automatization by default or the personalized option. To familiarize users with this strategy, the package provides a simulated dataset of industrial plastic bags fabrication.

Keywords: batch production, control, monitoring, DS-PC, package

Introduction

Nowadays, several industries rely on batch processing to yield final products. High-quality products are commonly described by quality characteristics (variables), each of which must be controlled within specifications to keep customer satisfaction and to describe the process performance as the batch progresses (Lewis, 2014). In this sense, statistical techniques are mandatory in control and monitoring of industrial processes, involving surveillance of correlated quality-process characteristics through control charts and other graphical methods (Bersimis et al., 2007).

First contributions to batch control and monitoring were mostly grounded on Multiway Partial Least Squares (MPLS) and Multiway Principal Component Analysis (MPCA) (Kourti et al., 1995; Nomikos & MacGregor, 1994, 1995). From that moment on, several strategies and methods have been developed in order to meet statistical assumptions, typical in batch production models (Lewis, 2014).

Alternative ways included Structuration des tableaux à trois indices de la Statistique (STATIS), Parallel Factors Analysis (PARAFAC), Tucker3, Multiset

Canonical Correlation Analysis (MCCA), Multiway Independent Component Analysis (MICA), Multiway Slow Features Analysis (MSFA), and Parallel Coordinates (Dunja et al., 2012; Escoufier, 1987; Filho & Luna, 2015; Harshman, 1970; Hyvärinen & Oja, 2000; Inselberg & Dimsdale, 1990; Jiang et al., 2018; Louwerse & Smilde, 2000; Meng et al., 2003; Parra, 2018; Tucker, 1966; Wang et al., 2017; Wiskott & Sejnowski, 2002; Zhang et al., 2017).

Some recent strategies even contemplate combining more than one technique to optimize multivariate batch process control under certain conditions, for instance, DS-PC approach (Dual STATIS-Parallel Coordinates). This nonparametric quality control strategy based on control charts enables off-line monitoring of batch processes using Dual STATIS and bagplots for control regions. A complementary analysis is developed with parallel coordinate plotting to examine tendencies within out-of-control batches. In this sense, this combined strategy brings on a variable-wise analysis, leading to support the visual interpretation of out-of-control signals (Ramos-Barberán et al., 2018).

Although the DS-PC strategy demand data preprocessing along with several calculations that are not particularly complex, its multi-step approach in which data set has been collected and cross-referenced in a series of steps that are fairly easy to mishandle, susceptible to mistakes and tough to replicate.

To facilitate DS-PC strategy application to any practical case, a package for the R environment (DSPC) was developed to automate the process and programmatically provide multivariate control charts and descriptive graphics for batch control and monitoring, without the uncertainty for miscalculation, and with the convenience and speed that computation provides. Then, industries will be able to determine if quality specifications are met, resulting in cost and time savings.

Methodology

The package DSPC was written for R, an open source software and programmatic environment for statistics and graphics. The program runs on the most popular computer platforms including Windows, MacOS, and UNIX. Since its inception, R has been used by programmers, scientists, practitioners, and code developers to create packages that guarantee reproducible code and results (Wickham, 2015). These packages run customized statistical functions, generate map and graphics, and allow researchers to import and export from large data sets in the public domain, among other uses. To date, many thousands of these packages have been developed in virtually all scientific fields and disciplines (Smith, 2017).

This package was elaborated as a generalization from the DS-PC strategy published by Ramos-Barberán et al. (2018). Development of the DSPC package followed coding and compilation guidelines outlined by Kim et al. (2018) and Leisch (2009). Functions and metadata files for DSPC were created and deposited on Git Hub, after a substantial testing period that started from 2018. Students at ESPOL

University were required the use of the package as part of an undergraduate class, and their experiences were used to troubleshoot and debug code.

Results

In this section, the main functions and outputs of the R package DSPC are presented. The functions of this package allow to perform the DS-PC strategy both in a step-by-step fashion and automatically. Table 1 shows a brief description for every function.

Table 1: Functions available in DSPC package

Function	Description
TabFactor	Create a Factor from a vector
SeqFactor	Create sequences over a Factor
TableObject	Create a Tables Object ready for Dual STATIS analysis
PreprocessTobj	Suggested preprocessing of the Table Object
DualSTATIS	Dual STATIS analysis for reference data
DualSTATISprojection	Dual STATIS projection of new tables
GenBagplots	Bagplots computing for projected tables. The compute.bagplot function contained in the aplpack package is used.
Parcoord2	Parallel Coordinates Plot. A modification of parcoord function from MASS package
AutoProcessing	Automatic processing of data using Dual STATIS
AutoPlotting	Automatic plotting of Dual STATIS results and parallel coordinates

Source: Self Made

Once the package is downloaded from the GitHub repository and ready to use in the R environment, it is possible to call these functions to perform analysis of batch processes with the DS-PC approach (Ramos-Barberán, 2020).

Data organization

Original data may be organized in separated multivariate tables with the same variables. Reference and additional batches data must be stacked separately, as shown in Figure 1.

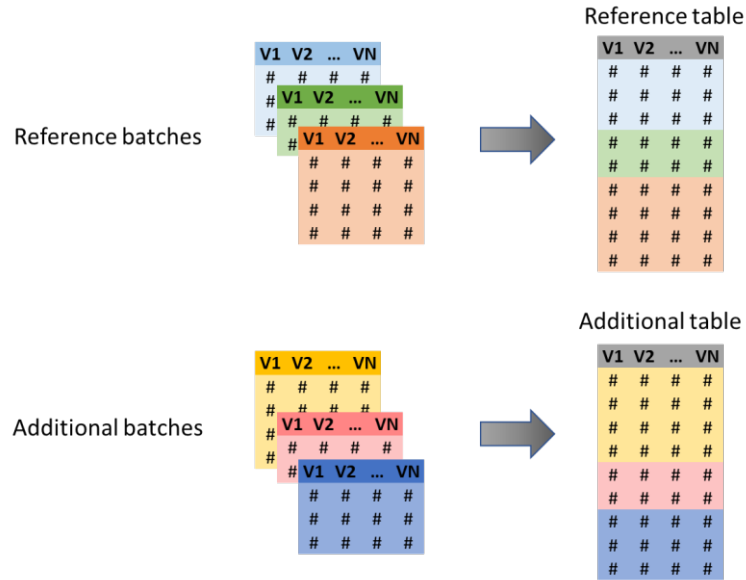


Figure 1: Data organization. Data from batches are organized in multiple tables for reference and additional testing, then, stacked into two matrices to load them properly to R.

Source: Self Made

As commented before, two main procedures can be considered: the suggested automatization and the personalized option. Despite the option preferred, loading the data tables into the R environment is required.

Suggested use

When a researcher is not familiar with the use of R, it may be overwhelming the amount of computational knowledge needed for the application of any robust method to the data available. If that is the case, this package considers a suggested option which is nearly automatic. To do so, follow these four steps:

1. *Prepare.* The reference data table must be contained in a csv file with a Factor-ObservationNames-variables structure, as appears in Figure 2.

Factor	ObsNames	Var1	Var2	...	VarN
A	SpecXY	###	###	...	###
A	SpecAN	###	###	...	###
A	SpecKN	###	###	...	###
A	SpecRS	###	###	...	###
B	SpecMR	###	###	...	###
B	SpecHJ	###	###	...	###

Figure 2: Data structure for DSPC package. The structure is organized by a factor to split the tables stacked, unique names for every observation, and data from every variable.

Source: Self Made

An additional data table, if required for monitoring scheme, must have the same Factor-ObsNames-variables structure.

2. *Load.* The reference table must be loaded into the R environment, then, the factor and names of observations should be extracted from this table. This loading step also must be done for the Additional table, if needed.

Supposing that Reftable.csv and AddTable.csv are prepared files with the structure shown in step 1 and placed in the current working directory, they can be loaded as tables and factors using the following script.

```
Reference = read.table(file = "Reftable.csv", header = TRUE, sep = ",")
Additional = read.table(file = "Addtable.csv", header = TRUE, sep = ",")

RefFactor = factor(Reference[,1])
RefTable = as.data.frame(Reference[,c(-1,-2)])
rownames(RefTable) = Reference[,2]
AddFactor = factor(Additional[,1])
AddTable = as.data.frame(Additional[,c(-1,-2)])
rownames(AddTable) = Additional[,2]
```

The separator character used in the function read.table may change. If the loaded table has not the correct format, it can be necessary to used ";" instead of ",".

3. *Compute.* Performing of Dual STATIS analysis is achieved using the AutoProcessing function in the pilot phase (just reference) or the monitoring phase (additional batches are considered).

```
#Pilot phase
res_Ref = AutoProcessing( RefTable, RefFactor )
#Monitoring phase
res_Ref = AutoProcessing( RefTable, RefFactor, AddTable, AddFactor)
```

4. *Plot*. The Interstructure, Intrastructure and Parallel Coordinates plots must be drawn using the AutoPlotting function and the Dual STATIS results derived from the step 3.

```
#Analysis mode
AutoPlotting( REF = res_Ref$REF, BPS = res_Ref$BPS )
#Monitoring mode
AutoPlotting( REF = res_RefAdd$REF,
              ADD = res_RefAdd$ADD,
              BPS = res_Ref$BPS )
```

Personalized use

When the data is organized as a Reftable and Addtable, both can be loaded to the R environment, and then use the available functions of the package to compute the elements according to the following:

1. If RefFactor is not already loaded, it should be created from Reftable using the TabFactor function.
 2. Optionally, if not available, names can be easily assigned to every observation in Reftable using the SeqFactor function on RefFactor.
 3. Creating of a table object (Tobj) using the TableObject function with the items Reftable and RefFactor.
 4. The preprocessing of matrices contained in Tobj\$Xk_data is suggested via scaling, centering and normalizing. The PreprocessTobj function allows to perform preprocessing on Tobj easily.
 5. Performing of Dual STATIS on Tobj via DualSTATIS function to generate a Dual STATIS results (Dsr). At this point, the Dsr list contains enough information to perform Interstructure and Intrastructure analysis. Additionally, the parcoord2 function allows to visualize all variables using the Tobj\$Original_data matrix.
 6. Steps 1 to 4 are repeated to obtain a table object (AddTobj) from Addtable. Then, use the DualSTATISprojection function on this AddTobj, taking Tobj into account, to obtain a Dual STATIS results list (AddDsr).
 7. GenBagplots function is used to create the control regions involved in the monitoring of additional batches. This completes the required elements for plotting of batches projections, complementing with parcoord2 function.
- A summarization of this scheme is presented in Figure 3.

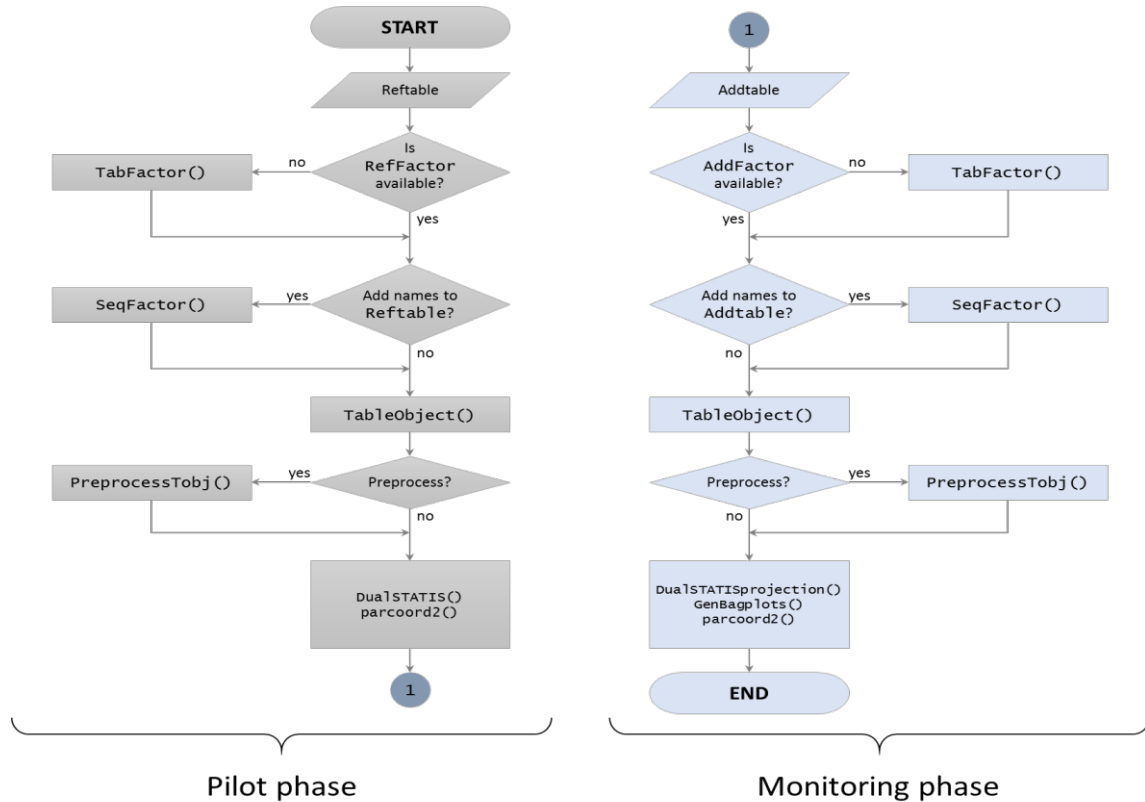


Figure 3: Personalized use of DSPC package. The flow chart shows the computation process for applying the DS-PC approach using the functions contained in the DSPC package. Source: Self Made

Available data

A simulated dataset of industrial plastic bags fabrication is available in this package. Every batch is represented by a table constituted by 50 observations of 3 process variables. A factor is associated to each data table. This dataset is coded as PlasticBags, a list containing 200 reference batches stacked in PlasticBags\$Ref\$data and 8 additional testing batches stacked in PlasticBags\$Add\$data. The testing batches are conformed by a normal batch and seven anomalous batches, affected by shifts in mean, standard deviation, and correlation shifts (Ramos-Barberán et al., 2018).

Outputs

Considering the plastic bag data, it is possible to easily generate all the plots associated, using the following script:

```
data( PlasticBags )
REftab = PlasticBags$Ref$data[1:1500,]
ADDtab = PlasticBags$Add$data
REffac = PlasticBags$Ref$factor[1:1500]
```



```
ADDfac = PlasticBags$Add$factor
DSPCres = AutoProcessing(REFtab, REFac, ADDtab, ADDfac)
AutoPlotting( DSPCres$REF, DSPCres$ADD, DSPCres$BPS )
```

Once the code has run, a set of plots is presented as result. These plots are collected in Figures 4 and 5. On the other hand, if the personalized option is chosen, the characteristics of the graphs can be modified, as it can be seen in Ramos-Barberán et al. (2018), section 3.1 Illustrative example.

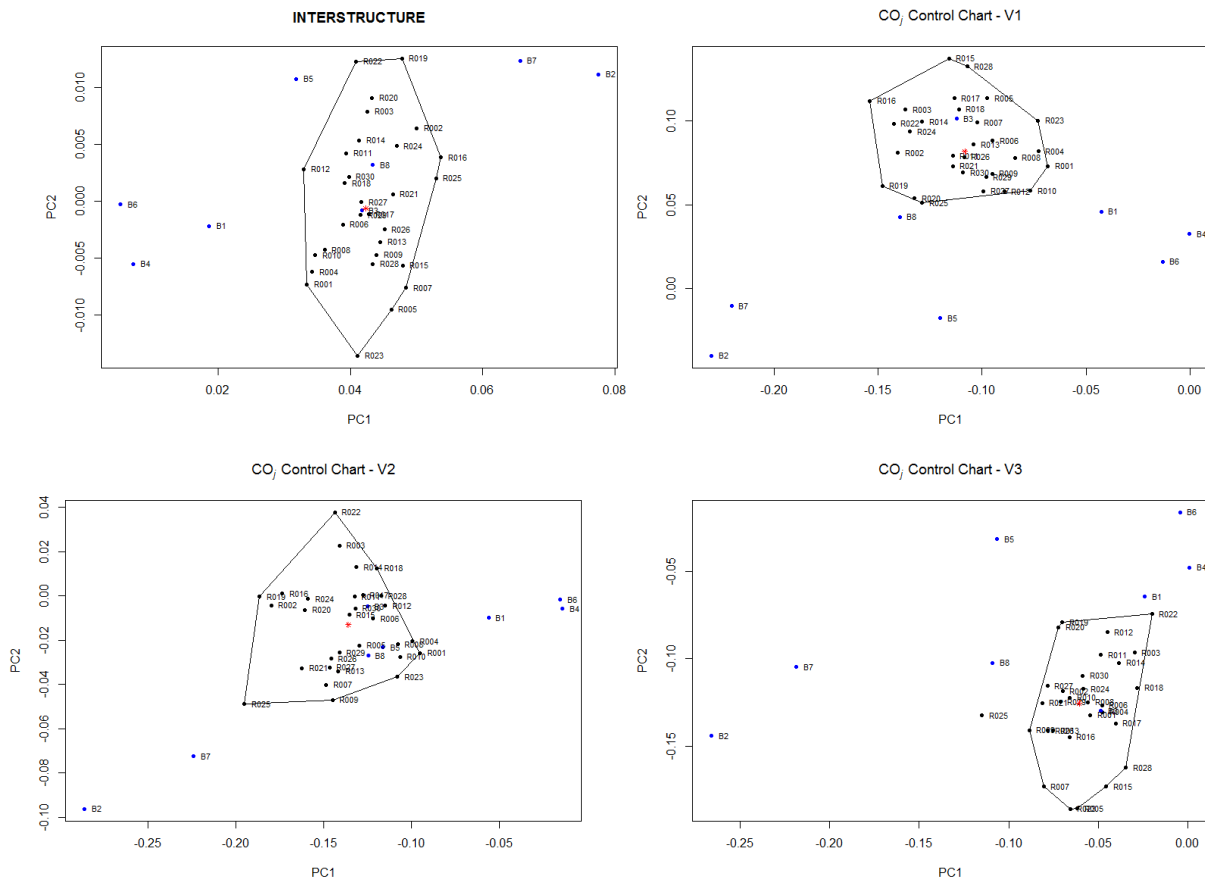


Figure 4. Control Charts created using the DSPC package. Graphs show the Interstructure (Left top) and Intrastructure Control Charts created using data and functions contained in the DSPC package.

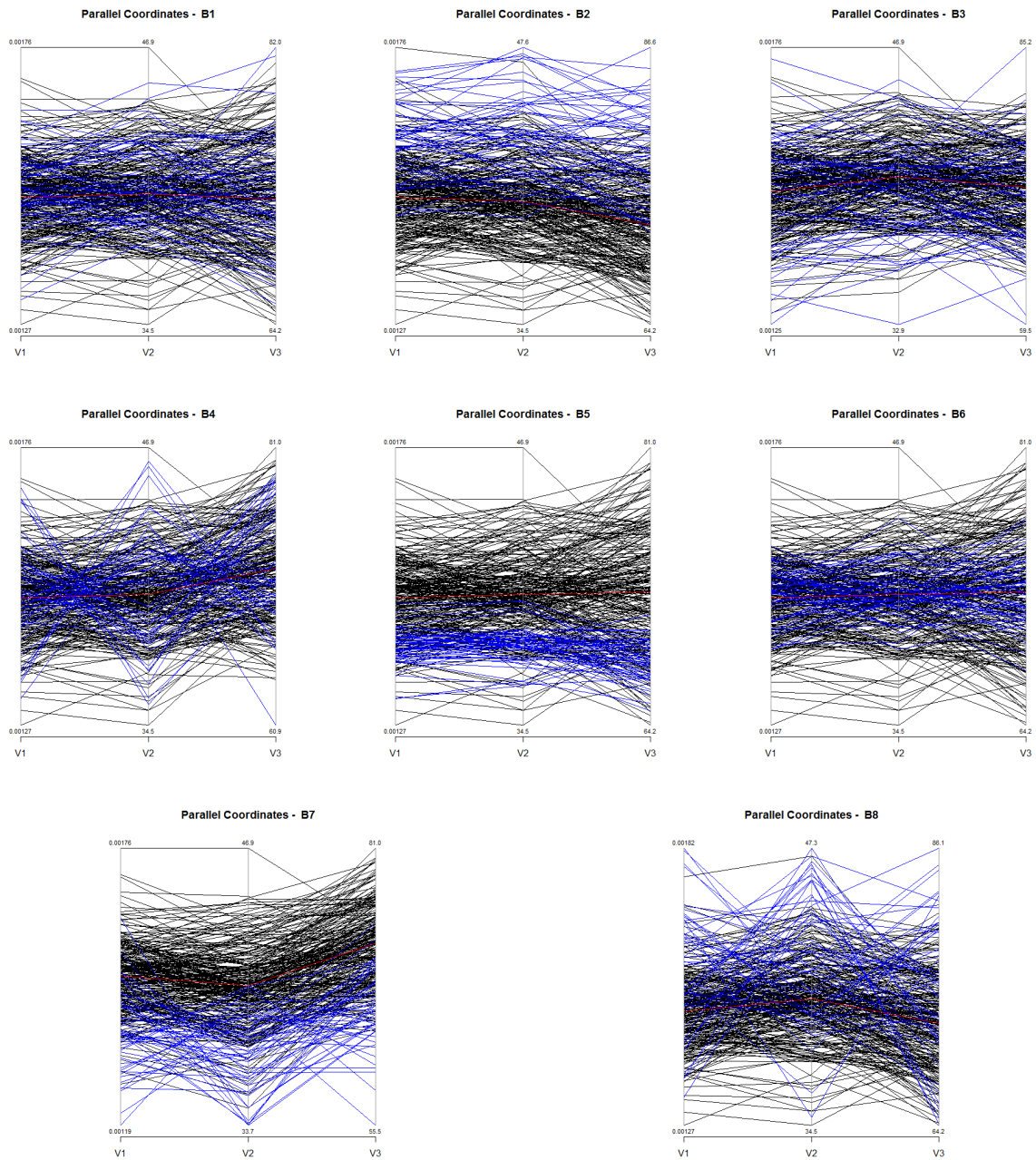


Figure 5. Parallel Coordinates plots created using the DSPC package. Graphs show Parallel Coordinates of reference data (black) and testing batches (blue), each of them, created using data and functions contained in the DSPC package.

Discussion

With the release of DSPC package, the authors hope that researchers will not only benefit from its practicality, but also, that they will explore a tool that conceptualizes Dual STATIS and Parallel Coordinates strategies under the scope of multivariate statistical process control and monitoring. Since these strategies have individual packages already uploaded to R repository in other fields applications (ade4, MExPosition, MASS), the implementation of this package would address a better comprehension of its combined potential for batch production application (Chessel et al., 2004; Chin Fatt et al., 2013; Dray et al., 2007; Venables & Ripley, 2002). It can also aid in the adoption of uniform historical records that can be used for monitoring across batches, variables, and time.

One of the conveniences of open-source software is that the original code is freely and easily available, and may be modified as needed. Future work on this package could incorporate control charts to handle batches with missing values, which are common in practice, as well as, other data preprocessing options. While these can be easily incorporated into this package, we believe it is best to show practical applications in multivariate statistical process control before it should be coded. In this sense, it is our hope that future versions of this package can be upgraded with feedback from researchers around the world.

Conclusions

To sum up, DSPC package is an easy time-saving graphical framework for scientists and practitioners to control and monitor any batch process data through DS-PC approach. It implements Dual STATIS computation for pilot and monitoring phases and offers graphic displays to detect unusual behavior during the production through IS (Interstructure) and CO (Intrastructure) control charts besides Parallel Coordinates outputs.

References

- Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate Statistical Process Control Charts: An Overview. *Quality and Reliability Engineering International*, 23(5), 517–543. <https://doi.org/10.1002/qre.829>
- Chessel, D., Dufour, A., & Thioulouse, J. (2004). The ade4 Package – I: One-Table Methods. *R News*, 4(1), 5–10.
- Chin Fatt, C., Beaton, D., & Abdi, H. (2013). *Package MExPosition*. <http://www2.uaem.mx/r-mirror/web/packages/MExPosition/MExPosition.pdf>
- Dray, S., Dufour, A., & Chessel, D. (2007). The ade4 Package – II: Two-Table and K-Table Methods. *R News*, 7(2), 47–52.
- Dunia, R., Edgar, T., & Nixon, M. (2012). Process Monitoring Using Principal Components in Parallel Coordinates. *American Institute of Chemical Engineers Journal*, 59(2), 1–12. <https://doi.org/10.1002/aic.13846>
- Escoufier, Y. (1987). Three-Mode Data Analysis: The STATIS Method. In B. Fichet & C. Lauro (Eds.), *Methods for Multidimensional Data Analysis* (pp. 259–272). ECAS.
- Filho, D. M., & Luna, L. P. (2015). Multivariate quality control of batch processes using STATIS. *International Journal of Advanced Manufacturing Technology*, 82(5–8), 867–875. <https://doi.org/10.1007/s00170-015-7428-0>
- Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16(10), 1–84.

- Hyvärinen, A., & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5), 411-430. [https://doi.org/10.1016/S0893-6080\(00\)00026-5](https://doi.org/10.1016/S0893-6080(00)00026-5)
- Inselberg, A., & Dimsdale, B. (1990). Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. *Proceedings of the First IEEE Conference on Visualization*, 361-378. <http://dl.acm.org/citation.cfm?id=949531.949588>
- Jiang, Q., Gao, F., Yi, H., & Yan, X. (2018). Multivariate Statistical Monitoring of Key Operation Units of Batch Processes Based on Time-Slice CCA. *IEEE Transactions on Control Systems Technology*, 27(3), 1368-1375. <https://doi.org/10.1109/TCST.2018.2803071>
- Kim, I. S., Martin, P., McMurry, N., & Halterman, A. (2018). *Instructions for Creating Your Own R Package*. <http://web.mit.edu/insong/www/teaching/teaching.html>
- Kourti, T., Nomikos, P., & MacGregor, J. F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *Journal of Process Control*, 5(4), 277-284. [https://doi.org/10.1016/0959-1524\(95\)00019-M](https://doi.org/10.1016/0959-1524(95)00019-M)
- Leisch, F. (2009). Creating R package: A Tutorial. In P. Brito (Ed.), *Compstat 2008- Proceedings in Computational Statistics* (pp. 1-19). Physica Verlag.
- Lewis, D. (2014). Control Charts for Batch Processes. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons.
- Louwerse, D. J., & Smilde, A. K. (2000). Multivariate statistical process control of batch processes based on three-way models. *Chemical Engineering Science*, 55(7), 1225-1235. [https://doi.org/10.1016/S0009-2509\(99\)00408-X](https://doi.org/10.1016/S0009-2509(99)00408-X)
- Meng, X., Morris, A. J., & Martin, E. B. (2003). On-line monitoring of batch processes using a PARAFAC representation. *Journal of Chemometrics*, 17(1), 65-81. <https://doi.org/10.1002/cem.776>
- Nomikos, P., & MacGregor, J. F. (1994). Monitoring Batch Processes Using Multiway Principal Component Analysis. *AIChE Journal*, 40(8), 1361-1375. <https://doi.org/10.1002/aic.690400809>
- Nomikos, P., & MacGregor, J. F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics*, 37, 41-59. <https://doi.org/10.1080/00401706.1995.10485888>
- Parra, L. C. (2018). *Multi-set Canonical Correlation Analysis simply explained*. *Nips*. <http://arxiv.org/abs/1802.03759>
- Ramos-Barberán, M. (2020). *Online repository for DSPC R package files data*. GitHub Repository. <https://github.com/mvramosb/DSPC>
- Ramos-Barberán, M., Hinojosa-Ramos, M. V., Ascencio-Moreno, J., Vera, F., Ruiz-Barzola, O., & Galindo-Villardón, M. P. (2018). Batch process control and monitoring: a Dual STATIS and Parallel Coordinates (DS-PC) approach. *Production and Manufacturing Research*, 6(1). <https://doi.org/10.1080/21693277.2018.1547228>
- Smith, D. (2017). *CRAN now has 10,000 R packages. Here's how to find the ones you need*. *Revolutions*. <https://blog.revolutionanalytics.com/2017/01/cran-10000.html>
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3), 279-311. <https://doi.org/10.1007/BF02289464>
- Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. <http://www.stats.ox.ac.uk/pub/MASS4/>



- Wang, Y., Jiang, Q., Li, B., & Cui, L. (2017). Joint-Individual Monitoring of Parallel-Running Batch Processes Based on MCCA. *IEEE Access*, 6, 13005–13014. <https://doi.org/10.1109/ACCESS.2017.2784097>
- Wickham, H. (2015). *R Packages*. O`Reilly Media. <http://r-pkgs.had.co.nz/>
- Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4), 715–770. <https://doi.org/10.1162/089976602317318938>
- Zhang, H., Tian, X., & Deng, X. (2017). Batch Process Monitoring Based on Multiway Global Preserving Kernel Slow Feature Analysis. *IEEE Access*, 5, 2696–2710. <https://doi.org/10.1109/ACCESS.2017.2672780>