MOdel based coNtrol framework for Site-wide OptimizatiON of data-intensive processes

D7.6 – Final Demonstrators Evaluation and Impact Report

Deliverable ID	**D7.6**
Deliverable Title	**Final Demonstrators Evaluation and Impact Report**
Work Package	**WP7 – Demonstration and Evaluation in the Aluminium and Plastics Domains**
Dissemination Level	**PUBLIC**
Version	**1.3**
Date	**2020-05-18**
Status	**Final version**
Lead Editor	**AP**
Main Contributors	**Vincent Maigron (AP), Massimo De Pieri (LCEN), Marc Fraysse (CAP), Marco Dias (GLN)**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 723650.

Published by the MONSOON Consortium

Document History

Version	Date	Author(s)	Description
0.1	2019-09-04	Vincent Maigron (AP)	First draft with table of content
0.2	2019-09-10	Massimo De Pieri (LCEN)	Contribution to chapter 3
0.3	2019-09-10	Marc Fraysse (CAP)	Add chapter 2
0.4	2019-09-17	Vincent Maigron André Augé (AP)	Adding AP contribution + merging all contributions
0.5	2019-09-24	Marco Dias (GLN)	Plastic domain contribution
1.0	2019-09-25	Vincent Maigron (AP)	Merging all contribution for final version
1.1	2020-03-09	André Augé (AP)	Contribution on chapter 3 and 4
1.2	2020-04-03	Marco Dias (GLN) Rosaria Rossini (LINKS)	Contribution on chapter 3 and 4
1.3	2020-05-18	André Augé (AP) Marco Dias (GLN)	Contribution on chapter 4

Version	Date	Reviewed by	Summary of comments
0.5	2019-09-25	Marco Dias (GLN)	Accepted with minor comments
0.5	2019-09-27	Ruben Schlutter (KIMW)	Accepted with minor comments

Table of Contents

Executive Summary

This document constitutes the Deliverable D7.6 – Final Demonstrators Evaluation and Impact Report of the MONSOON project (Grant Agreement No. 723650) and presents a study about the impact of dataflow simulation, the final results of defined KPI's, and their impact on each domain business cases.

1 Introduction

This document is the final evaluation of the KPI framework defined in D7.1 and D7.2 from task 7.1. This task is a part of WP7, which is dedicated to stablishing tools to evaluate the effectiveness of the application of data-driven optimization methodologies developed by data scientists along with other tasks of the project and validation of the presented methodologies.

This deliverable firstly presents a study on data flow simulation, as an accelerator for predictive function development. A literature study is developed, explaining the state of the art of this subject. Then the current and future impact are described.

The final analysis and associated calculations of the KPI framework defined and implemented during the project are also presented, according to the actual needs/requests, based on each domain (Aluminium and Plastic) business cases. The KPIs were categorized into different clusters as it was described on deliverable [RD.1].

1.1 Related documents

ID	Title	Reference	Version	Date
[RD.1]	Initial Evaluation Framework	D7.1	1.0	28-03-2018
[RD.2]	Final Evaluation Framework	D7.2	1.5	19-04-2019
[RD.3]	Final Life Cycle Management plugin	D5.8	0.5	01-07-2019
[RD.4]	Final Online and Deep Machine Learning Functions	D5.4	1.1	01-08-2019
[RD.5]	Initial Demonstrators Evaluation and Impact Report	D7.5	1.0	19-03-2019

2 Data Simulation impact

Data simulation has been studied for a potential integration in the MONSOON Datalab. The initial goal was to provide data scientists with additional data in the form of time series reproducing the behavior of the real plant data.

Such approach can provide multi benefits. Increasing the volume of data can allow the data scientist to develop, train and test machine learning algorithms, at least going as far as defining a model and its hyperparameters. Using simulated data can increasing the MONSOON platform's internal and external confidentiality, by not sharing unencrypted process data with anyone, only providing the associated simulated data. Missing data due to faulty sensors or external causes can also be simulated, providing the data scientist with more robust training datasets.

2.1 Literature study

This section presents the current state of the art regarding the simulation of random data, correlated datasets, time series and correlated time series. A study about the extraction of key features from real process data, needed to generate accurate synthetic data, as well as the analysis of correlation between real datasets and time series is presented. Finally, the interpolation and extrapolations methods that could allow computing the missing data from real data extractions is studied.

Most programming languages provide tools for generating random data according to a predefined distribution. This allows for an easy generation of random datasets following real life distributions. In the MONSOON project however, two additional difficulties make generating data more difficult. First, all data is acquired as timeseries. Second, multiple timeseries are used to extract meaningful information, and they are correlated in time and space.

2.1.1 Generating synthetic timeseries

Timeseries simulation is a widely studied field and several approaches allow generating timeseries following a defined global trend as well as reproducing cyclical events. For example, the Arima model takes as input a timeseries and produces as output the values the timeseries is expected to take in the future[1].

Several tools allow the user to create time series according to predefined parameters, such as a global tendency as well as seasonal trends and patterns[2].

2.1.2 Generating correlated datasets

In the case of non-temporal random datasets, accounting for the correlation between sets is generally achieved using covariance matrices[3]. These matrices list the correlation between all pairs of datasets and can be estimated using premade tools[4].

[1] Arima for simulating time series : https://pkg.robjhyndman.com/forecast/reference/simulate.ets.html
[2] Time series simulation tool : https://cran.r-project.org/web/packages/eesim/vignettes/eesim.html
[3] Using covariance matrices to simulate correlated data : https://en.wikipedia.org/wiki/Estimation_of_covariance_matrices

A potential approach for generating synthetic data that keep the real data's main characteristics is to estimate the covariance matrix from the real data and then generate new sets from this matrix.

Other more recent approaches aim at keeping more of the real data's characteristics by employing various mathematical tools[5] [6].

Promising novel approaches have been developed recently, such as the Synthetic Data Vault technique[7], that generates synthetic relational datasets keeping the most important characteristics of the original data. Results in this field seem promising, as this method has allowed several data scientist teams to develop machine learning algorithm solely based on the synthetic data that produced results as good as the ones developed using the real data.

2.1.3 Generating correlated timeseries

Generating correlated timeseries imparts additional challenges. Keeping Seasonal, periodical and punctual events in line with the real data while generating adequately correlated timeseries has made the research field slower to bear fruits.

Some methods rely on strong assumptions to obtain partial results[8], but, to our knowledge, none come close to proving their generated data is close enough to the real timeseries to be used as a base for data science studies.

2.1.4 Extracting key features and correlation relations from real data

An interesting approach to generating adequate timeseries from real data could be starting by extracting key features from data and only then try and find how to produce data from these features. This field has known some interesting advances in recent years, with approaches extracting new information from multivariate time series[9].

2.1.5 Interpolation and extrapolation

Instead of generating all new timeseries, it can be useful to try and fill the gaps in production data, or even to infer new values at a smaller timestep than originally used.

---

[5] Simulating new multicriteria data from a given data set : https://eric.univ-lyon2.fr/~arolland/publis/DA2PL-CUGLIARI-ROLLAND_v5.pdf
[6] Simulation of Correlated Data with Multiple Variable Types Including Continuous and Count Mixture Distributions : https://cran.r-project.org/web/packages/SimCorrMix/index.html
[7] Simulating correlated relationnal data based on real data : https://dai.lids.mit.edu/wp-content/uploads/2018/03/SDV.pdf
[8] A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence https://www.itia.ntua.gr/en/docinfo/1488/
[9] Mining Novel Multivariate Relationships in Time Series Data Using Correlation Networks : https://arxiv.org/pdf/1810.02950.pdf

This is widely achieved using interpolation, but one should remain aware of the limitations of such approaches[10]. Extrapolation methods can generate interesting new data posterior to the real extracted production data, but most methods can sometimes be widely imprecise.

Some recent approaches have achieved interesting results in generating missing data using machine learning[11]. However, they too can sometimes lead to inacurate estimations.

2.2 Current and future impact of data simulation

Many, if not most of the methods described state-of-the-art, fall short of providing insurance that the generated data will be close enough to the real data. The MONSOON project workflow would not benefit from the addition of simulated data that does not adequately enough represent the ground truth.

The actual impact of studying data simulation for the MONSOON Datalab has been the use of some restricted versions of the algorithms mentioned above. Some approaches have been definitively cut off, as their results were not accurate enough. Finally, some methods still could lead to interesting tools, but were not mature enough to be used during the MONSOON project.

2.2.1 Applied approaches

Two main contributions have been made to the MONSOON Datalab workflow using data simulation tools. They are both introduced in the D5.4.

First, as several timeseries extracted from sensors are used to derive classification and prediction models, the datasets have been balanced to contain a similar number of “positives” and “negatives”. For example, when predicting equipment stoppages that occur only rarely, stoppage events have been simulated and included in the timeseries. This has been done using Scikit learn included tools, with custom settings.

Second, as some data presented high dimensionality, preventing efficient analysis by machine learning algorithms, tools have been used to reduce this dimensionality, effectively simulating a reduced dataset containing the essential information. For this purpose, Scikit learn’s Principal Component Analysis tool has been leveraged and customized.

2.2.2 Limited approaches

Many approaches introduced above cannot be included in the MONSOON project’s workflow. Indeed, they induce too much imprecision in the generated data, or do not suit the actual shape of the process data extracted.

Extrapolation methods can introduce quite high imprecision on the generated data and struggle to reproduce punctual events in timeseries. Hence, they were not included in the MONSOON Datalab.

---

[10] Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment (Mathieu Lepot , Jean-Baptiste Aubin and François H.L.R. Clemens)

[11] BRITS: Bidirectional Recurrent Imputation for Time Series : https://arxiv.org/pdf/1805.10572.pdf

Interpolation could be used to produce more fine-grained data, but it would not provide more information to data science algorithms, and thus would not enhance the datasets. Most interpolation methods struggle to fill data in long time gaps, and so they could not reliably provide missing sensor data.

A way to combine Arima time series simulation and standard covariance matrices used in random datasets by correlating Arima models' parameters instead of the variables directly has been envisioned but lead to no interesting results.

The opposite approach, in which we first generate properly correlated random datasets and then induce seasonal, cyclical and punctual events has not led to any promising result either.

As a last resort, one could ignore all correlation between the synthetic timeseries and only employ an array of Arima simulations. It would however lead to unacceptable imprecisions, as can be seen in the results of the article "Simulating new multicriteria data from a given data set[12].

Creating synthetic data can be achieved by simply adding well chosen noise. It however does not provide different enough data to be used as a mean of confidentiality and does not serve to better train machine learning algorithm either.

Promising approaches in related fields could not properly be applied, such as the method described in "The Synthetic data vault"[13], in which the generated data takes the form of relational data. Same goes for Using linear combination of independent variables[14], that would provide datasets too different from the ground truth to be usable**.**

2.2.3 Future approaches

Some approaches could have provided interesting results in the Datalab but were not mature enough to be used during the project.

The approach described in "Mining Novel Multivariate Relationships in Time Series Data Using Correlation Networks"[15] gives promising results in extracting key features and correlation relations between time series. There is however no robust way to construct all new timeseries based on these extractions yet.

The BRITS[16] algorithm provides a promising alternative to interpolation for retrieving lost sensor data. It is but a first step in the right direction, as results are still not robust enough and involve hard to customize machine learning algorithms.

---

[12] Simulating new multicriteria data from a given data set : https://eric.univ-lyon2.fr/~arolland/publis/DA2PL-CUGLIARI-ROLLAND_v5.pdf

[13] Simulating correlated relationnal data based on real data : https://dai.lids.mit.edu/wp-content/uploads/2018/03/SDV.pdf

[14] Spatial-temporal Causal Modeling for Climate Change Attribution: https://www.researchgate.net/profile/Alexandru_Niculescu-Mizil/publication/221654446_Spatial-temporal_causal_modeling_for_climate_change_attribution/links/02e7e52249e5ea256f000000.pdf

[15] Mining Novel Multivariate Relationships in Time Series Data Using Correlation Networks: https://arxiv.org/pdf/1810.02950.pdf

[16] BRITS: Bidirectional Recurrent Imputation for Time Series: https://arxiv.org/pdf/1805.10572.pdf

Finally, there exist some algorithms[17] that directly try to generate synthetic timeseries reproducing interesting behaviours in the real data. They are however not complete enough yet and may miss some important characteristics present in the actual process. Those could impart unexpected difficulties in the development of machine learning tools and as such cannot yet be included in the MONSOON platform.

[17] A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence: https://www.itia.ntua.gr/en/docinfo/1488/

3 Domain KPIs Evaluation Framework

For both the industrial domains investigated by MONSOON the evaluation framework depicted in [RD.2] applies. Even though complete description of the rationale behind this tool is out of the scope of this deliverable, a brief recap is reported here to ease the understanding of this document.
Evaluation framework composes of several layers which can be applied to any of the industrial processes where the MONSOON platform is applied. The layers are the following:

1. ***Domain-specific KPIs***: this class includes all the indicators that are used to quantify the onsite effectiveness of the optimization functions. As a general consideration, these KPIs can't be cross-sectorial due to their high connection with the specific use cases and cannot be transferred to other domains (unless in case of particular conditions of technological similarity);

2. ***Environmental KPIs***: this class includes all the indicators employed to quantify specific environmental impacts. These figures are used to quantify both the local and the global footprint related to the investigated process; typical examples of environmental KPIs are the total greenhouse gases emission, the use of resources, water consumption. These KPIs present high cross-sectoriality, as they can be used to evaluate the environmental impact of a wide range of manufacturing processes;

3. ***Replicability KPIs***: this class includes aspects which are not related to the process but to the MONSOON platform itself. As one of the goals of the project is to provide a toolkit which can be adopted to boost efficiency in the process industry, the easiness of deployment of the platform should be carefully evaluated to ensure a market breakthrough.

The first class is specific of any investigated use case and no general considerations can be performed. Environmental and Replicability KPIs are cross-sectorial for their own nature and can be applied to evaluate the MONSOON effectiveness to several different industrial applications. In the following subsections a description of the environmental and replicability KPIs is reported and most relevant features are described.

3.1 Environmental KPIs

Environmental indicators are used to evaluate the eco-efficiency of a process. The larger the number of adopted indicators, the wider the perspective from which the sustainability pillar is evaluated. KPIs listed in the following table have been validated and selected during T7.1 activities. The origin is typically the LCA plugin developed within WP5 tasks: for further clarifications about this component and its role into the MONSOON platform, please refer to D5.8.

Table 1 - Evaluation framework: environmental KPIs

KPI name	Unit	Origin	Class	Description
Global Warming	kg CO_2 equivalent	LC management plugin	Environmental	Total amount of equivalent greenhouse gases generated by the investigated process

KPI name	Unit	Origin	Class	Description
Primary Energy Consumption	MJ equivalent	LC management plugin	Environmental	Total amount of primary energy required to manufacture the investigated product
Direct Energy Consumption	MJ equivalent	LC management plugin	Environmental	Total amount of energy (electric and thermal) directly consumed by the investigated process
Electricity consumption	MJ	LC management plugin	Environmental	Total amount of electricity directly consumed by the investigated process
Raw material consumption	kg	LC management plugin	Environmental	Total amount of material required to manufacture a unit of valuable product
Recycled content	%	LC management plugin	Environmental - process	Percentage of recycled material in the investigated product
Water Consumption	l	LC management plugin	Environmental	Total amount of water required to manufacture the investigated product
Waste to landfill	kg	LC management plugin	Environmental	Total amount of waste originated from the process which manufactures the investigated product and sent to landfill
Waste to recycling	kg	LC management plugin	Environmental	Total amount of waste originated from the process which manufactures the investigated product and sent to recycling
Process yield	%	LC management plugin	Process	Ratio between the valuable output of the investigated process and the total input material

3.2 Replicability and scalability KPIs

Replicability plays a key role in ensuring industrial competitiveness of the MONSOON platform. For this reason, a set of indicators describing how to estimate the replicability of the service has been defined and proposed by WP7 and shown in Table 2. Note that the specified KPIs are related mainly to the Real-time operational platform. The Data Lab platform provides generic tools, which can be applied on the broad list of data analytical problems without the modification of the architecture or introduction of the new software components. The exceptions in Data Lab include the implementation of the new data analytical algorithms, e.g. for the research purposes. This case is supported by the application development interface (API) provided by the open-source Python data analytical stack adopted in the MONSOON platform. Otherwise, the development of new predictive functions for the new applications is fully supported without the platform modification.

The new platform components are required mainly for the data integration layer in the Real-time operation platform, where the deployment of the platform to the new site can require the implementation of the proprietary data connectors for the specific data sources. In the MONSOON, the data integration layer is implemented using the Apache Nifi technology, which already provides the implementation for a wide range of standard data protocols and formats. The Apache Nifi also has the modular architecture and provides API for simple implementation of the new proprietary connectors and data transformations.

The replicability is also supported by the packaging of the platform components using the Docker technology, which reduces the configuration tasks and allow to easily deploy customized configuration of the software or update specific component without the influencing the rest of the platform.

Table 2 - Replicability KPIs

Indicator	Unit	Meaning
Number of configuration changes	amount	Number of changes in configuration setup required to adapt MONSOON to another domain different to the pilot ones. A change is defined as every kind of modification in whatever infrastructural component
Estimated Effort for update	low-medium-high	An estimation of the effort required to update a connector to retrieve different data from different machines or data acquisition systems
Infrastructure adaptation	amount	Amount of new components which need to be developed to ensure the proper operation of MONSOON from infrastructural point of view

4 Evaluation Results

4.1 Aluminium Domain

4.1.1 Evaluation of the results and improvements

Based on the green anode production and based on the evaluation framework defined in D7.1, here the following results are presented:

Table 3 - Evaluation Framework applied to aluminium domain, (average green anode production, Aluminium Dunkerque plant). Green lines represent the base layer; blue lines contain domain specific KPIs.

KPI name	Reference flow	Unit	KPI value 2017	KPI value 2019	Comments
Global Warming Potential	Anode	kg CO_2 equivalent	566	566	
Primary Energy Consumption	Anode	MJ equivalent	60,480	60,480	
Direct Energy Consumption	Anode	MJ equivalent	565	565	
Electricity consumption[18]	Anode	MJ	55	55	
Raw material consumption	Anode	kg	1082	1084	
Recycled content	Anode	%	26.1	26.1	
Water Consumption	Anode	l	5,400	5,400	
Waste to landfill	Anode	kg	0.2	0.2	
Waste to recycling	Anode	kg	0.3	0.3	
Product Circularity Index	Anode	%	N/A	N/A	This KPI cannot be computed for intermediate products
Anode quality	Batch	%	8.5	5	Percentage of 30 minutes periods with lower anode density with a given threshold. By definition threshold changes every day in order to

[18] During iteration 1, only mixer electricity has been considered in the analysis

					have 5% of bad periods
Anode rejection rate	Anode	%	N/A	N/A	This KPI has not be computed during iteration 1[19]
Acidification Potential	Anode	kg SO_2 equivalent	9	9	
Natural gas consumption	Anode	MJ	510	510	

The focus has been put on the green anode production, and the main KPI associated is green anode rejection rate.

For iteration 1, AD was working on batches of anodes produced per 30 minutes' periods. A period is considered of lower quality if at least one anode produced has a density below 1.62. The KPI is therefore the number of such lower quality periods.

The anode quality predictive function is thus being further adjusted considering plant performance. Moreover, a performance protocol has been also put in place to promote a more extended evaluation of the function, also after the end of the project. The protocol is defined and describe hereafter. During a shift of stabilized production (that means after at least one shift of production without stoppage)

- Check of repeatability (repeatability of the recommendations): 3 consecutive thirty minutes periods classified improvable without doing anything except normal exploitation of the paste plant.
- Change of process parameters according to predictive function recommendations.
- Comparison: After stabilization of the process after the change, comparison of median and average densities between the 3 improvable periods and the 3 periods following the modification. Check that these periods are good.
- Comparison of the density reachable given by the predictive function with reality.
- Check of reproducibility: the former process is repeated 3 times.

The objective of the performance protocol is the first step of the on-going extended assessment of the impact of the anode quality predictive function. If the results of the performance protocol tests are good, Liberty Aluminium Dunkerque will have to generalize the use of the recommendations of the anode quality predictive function and to change the operating procedures and to train the whole team of the paste plant (using the same training material used to train the team members involved in the project demonstration).

The Life Cycle Assessment dashboard running at the paste plant is a tool for monitoring in real time (by shift of 8 hours). Among KPI are the use and re-use of raw carbon materials on one hand, energy consumption on the other hand. LCA is definitely a useful dashboard for the paste plant process engineer. Depending on the efficiency of the predictive maintenance function, change of scheduling of maintenance would be done.

[19] Aluminium Dunkerque fixed the upper limit value for this indicator at 2.5%

4.1.2 Aluminium Domain Impacts

The impacts of the anode quality predictive function are firstly on energy and secondly on use of raw materials. The prediction of the bath height would make possible to control the control of this quantity. Thus, we would avoid cases where the bath height is too low, leading, among other things, to anode or mudding effects (because the volume of bath for the dissolution of alumina would be lower). The anode effects cause over-consumption of electricity when they occur and a decrease in Faraday efficiency. The mudding phenomena add an additional resistance in the pots and create instabilities, therefore, once again, increase in electrical consumption and loss of Faraday efficiency. If the bath height is too high, there is a risk of the logs dissolving and as a consequence pollution of the aluminum produced by iron.

4.1.2.1 Energy saving

In details, the predictive anode quality function helps to put under control the green anode production process. The consequences are a better robustness of the process and a decrease of deviations of anode quality. The cost of a crisis in the electrolysis department, consequence of low anode quality is 10 M€ mainly due to loss of efficiency on energy use. That means that the impact of anode quality predictive function is an improvement on energy consumption on pots. There are one such a crisis every four years in average and the gain due to the anode quality predictive function is estimated to be ***300 k€/year*** or **12%** of total crisis costs.

Despite major changes during the project, data scientists of the consortium (CERTH team) have worked on predictive functions on electrolysis process (cf. WP5). Few results have been achieved, also addressing some difficulties emerged during the project. Because of these ones, it was not possible to put in full operation the solution in Liberty Aluminium Dunkerque (as instead done for the anode quality predictive function); however, an evaluation has been performed offline with real data collected from the same plant.

4.1.2.2 Use of raw materials

A consequence of the improvement of anode quality due to anode quality predictive function is a density increase. That means that the weight of every anode is higher. On a chemical point of view (without any parasite reactions), the aluminium production is proportional to anode consumption. For a given anode, the weightier is the anode, the larger is the weight of aluminium produced with this anode. So, more aluminium is produced during the life cycle of the anode on the pot. That avoids producing extra anodes in case of electrolysis intensity augmentation. The gain for Liberty Aluminium Dunkerque should be ***150k€/year***. The gain is in better use of raw materials, coke and pitch.

4.1.2.3 CO_2 emissions

CO2 emissions in aluminium production arise from 2 main sources: anode production and aluminium electrolysis. While the latter has a significantly higher environmental footprint, due to the high amount of energy consumed by the process and the energy-intensive bauxite refining process to obtain alumina (the basic substance for the electrolysis), anode production might have a significant, indirect contribution.

An anode with bad quality can affect the electrolysis process, compromising the whole batch of aluminium produced in the pot. In this case, the environmental (and economic) footprint is significantly heavy, as no valuable output is produced but several amounts of energy and material are consumed, thus leading to a loss in terms of resources and money.

Low-quality anodes can potentially affect bath height, leading again to inefficiencies in terms of current intensity and distribution inside the electrolytic pot.
From direct impact point of view, anodes are produced in the paste plant and have an average weight of 1 ton. According to field-specific data, elaborated with the LCA plugin described in D5.7 and 5.8, about 550 kg of equivalent CO_2 are generated to produce a single anode. If the anode has bad properties in terms of density or any other quality issue arises, the anode must be rejected. This means the material is recovered, but the process shall take place again, consuming additional energy. From site-specific data, around 2,5% of total anodes produced at Dunkerque were rejected. This corresponds to 3 800 anodes that generated greenhouse gases but were then re-processed due to quality problems. Considering the average from LCA plugin, this implies a Greenhouse gas emission around 2100 tons of CO_2. The MONSOON quality predictive function will help to reduce the rejected rate toward 0. As a consequence, the impact on CO_2 emission will be around 2100 tons per year.

4.1.2.4 Process optimization

The MONSOON predictive function anode quality is contributing for the improvement of anode density. The recommendations of the predictive function give information to Paste plant process team on both actuators to change and the value for the associated new set point. The MONSOON predictive function anode quality is a support to the process team to take decision in real time. The automation of such a predictive function - in case of success - should be studied.

As the anodes are used in the electrolysis process, the MONSOON predictive function quality has an impact on global optimization of the plant. For example, as a consequence, as described above there is an impact on electrolysis process by reducing the number of crisis. In the absence of crisis during the electrolysis process, the aluminium's flow from electrolysis department to the cast house is optimized in order to improve delivery times to customers and more generally the customers' satisfaction.

In case of success of the MONSOON predictive function anode quality at Aluminium Dunkerque, it could be deployed within Rio Tinto paste plants in Canada (3) and could be also deployed within Rio Tinto's joint ventures depending on their own strategy, Europe (Netherlands), Canada and Australia. The MONSOON predictive function anode quality could have a worldwide impact on aluminium production of Rio Tinto smelters and joint ventures.
Also the Lifecycle Assessment dashboard has a positive impact by giving information in real time on environment data.

4.1.2.5 Other

The improvement of anode quality will also minimize the number of abnormal situations on pots. It will reduce the number of anode change out of schedule, with as consequence less exposure to deteriorated conditions for the operators and less emission.

4.2 Plastic Domain

4.2.1 Evaluation of the results

Based on the iteration 1 defined on D7.1 related to the coffee capsule production, related to the evaluation framework applied to plastic concerning the year 2017, were presented the following results, based on a new calculation between 2018 to 2019:

Table 4 - Evaluation Framework applied to plastic domain, (comparison of coffee capsules production between year 2017 and October 2018 to June 2019, GLN facilities). Green lines represent the base layer; blue lines contain domain-specific KPIs.

KPI name	Reference flow	Unit	KPI value 2017	KPI value 2019	Comments
Global Warming	kg of capsules	kg CO_2 equivalent	2,7	2,1	
Primary Energy Consumption	kg of capsules	MJ equivalent	87	79	-
Direct Energy Consumption	kg of capsules	MJ equivalent	8	3	-
Electricity consumption	kg of capsules	MJ	8	3	-
Raw material consumption	kg of capsules	kg	1,26	1,02	-
Recycled content	kg of capsules	%	0	0	-
Water Consumption	kg of capsules	l	18	15	-
Waste to landfill	kg of capsules	kg	-	-	-
Waste to recycling	kg of capsules	kg	0,26	0,02	-

KPI name	Reference flow	Unit	KPI value 2017	KPI value 2019	Comments
Process yield[20]	injection process	%	79	98	-
Rejection rate	kg of capsules	%	3	2	-

The presented results refer to the coffee capsule production previous sampling described on [RD.5] and the comparison to a new sample between October 2018 and June 2019.

In a first sight and comparing to the first sample, the new values reveal that improvements were achieved regarding the consumption of resources and the efficiency of the processes. However, part of this deviations is related to the factor that for the first sample, in 2017, plastic transportation to recycling site and electricity consumption for the chiller were considered; these inputs were not available for this new sample calculation, namely the Global Warming and the consumption of resources - Primary and Direct Energy Electricity, Raw Materials and Water. Nevertheless, the trend points to the positive evolution of the KPI's, based on the tendency of the Waste to Recycling, Process Yield and Rejection Rates.

The presented results validate the possibility for process improvements, concerning the raw material and energy consumptions, as anticipated in [RD.5]. In the next chapter, it will be commented and explained the relevant impacts obtained, based on these figures.

4.2.2 Plastic Domain Impacts

As written in the previous document [RD.5] for the plastic domain, the results reflect significant environment impacts once the product scope is based on plastic parts.

Before MONSOON the expectations of improvements, for the plastic domain, were pointed for the decrease of on-site material handling time (by 10%), the decrease of resources consumption (by 10 %), the decrease the global use of energy on-site (by 10%) and the decrease of the Green House Gases emissions by 10%.

As described in [RD.5], the usage of the plastic is a major concern because of mass consumption and the waste issues. The application of the MONSOON solution on this segment will contribute to upstream improvements, minimizing impacts on the final output.

Through MONSOON solution, we can visualize the plastic injection process impacts and identify the main actors and where to act. One example is the forecast of unexpected production stops and mold and machine maintenance prediction. With these tools, the management team can act directly on the production planning, and schedule stops for maintenance intervention, avoiding waste production and lack of deliveries to his costumer, according to the weekly supply contract.

[20] This significant deviation from 2017 to 2018-2019 is majorly related to issues on the molds, when the calculations were preformed; at the same conditions (from 2018-2019), the Process Yield for 2017 should have been a bit better.

More details on plastic domain impacts are reported below, mainly considering the global optimization based on anomaly detection for predictive maintenance and raw data (in real-time).

4.2.2.1 Use of raw materials

From the perspective of achieved results, GLN has driven a scenario of 12-15% improvement in terms of the direct costs savings: considering the two use cases we can assume an improvement of the rejection rates from 3 to 2% and an estimation of savings of ~€42k/year on consumption of the raw materials, based on 200kg/day reduction of plastic waste. These values were calculated according to the improvements on the detection of anomalies on both use cases – coffee capsule production and automotive plastic parts.

However, for the second business case it is expected to achieve some additional gains by fixing few remaining issues in the installation of in-mold sensors whose measurements will be exploited by the anomaly detection predictive function for the mold faults.

4.2.2.2 Energy saving and reduction of material handling time

In terms of indirect cost, namely the material handling time and energy, these factors were included in the global optimization tools, and GLN estimated a gain of 3% of global savings for the core business, translated in ~€210k/year. These estimations were performed based on the global consumption of energy of the unit, once we cannot segregate per process/machine and the material handling time involved in the injection process, and represents an estimation of ~12% decrease of global use of energy on-site) and ~15% decrease of material handling time.

4.2.2.3 Process optimization

For this approach, the MONSOON **predictive tool** is contributing for the improvement of the injection process in the coffee capsule production. Getting to know in advance if the cavities are going to stop, allows the production manager to handle the problem with time and take action: to apply a correction through process variables or to decide if the mold tool needs maintenance, avoiding the production of bad parts/waste and energy consumption.

In another vector, this tool also contributes for the global optimization, proving information for the analysis of the effectiveness of the process which allows the manager to have the full picture of the profitability of the production in real-time and evaluate if an action is needed.

The impacts of the tools are correlated to the gains of raw materials consumption and the improvement of the indirect costs, already described in the previous subsection.

Also, the Lifecycle Assessment dashboard had a positive impact promoting actions toward increased process efficiency. In fact, for the plastic domain, only an offline model was performed which gave a good perspective of the production area in terms of the environmental impacts. Nevertheless, concerning the energy consumption, the board identified the electrical energy as a factor to be controlled. Aligned to it, several energy analyzers were installed into the injection machines in the last months, in order to collect the real consumptions per machine. These analyzers are not yet connected to the MONSOON solution due to suppliers constrains, but overcoming this situation, GLN will be able to use the MONSOON LCA plugin, and in the next months associate the real-time data to the global optimization, providing the real scenario of the injection processes, allowing the evaluation of the environment KPI's, namely the energy consumption per plastic part.

4.2.2.4 CO_2 emissions

Concerning the Green House Gases emissions, the CO_2 savings have been computed starting from results obtained within WP5 work and documented in LCA-related deliverables [RD.2 and RD.3]. The analysis of material and energy flows in two different years allowed to compute the GHG impact for 1 kg of good parts ready for the market; these values were quantified in 2,7 kg of CO_2eq and 2,1 kg of CO_2eq respectively. In relative terms the reduction is about ~12% compared to the baseline scenario (2017 production).

This reduction is mainly due to the awareness brought from project to the shop floor and to the management teams, which become more sensitive to the ecologic footprint associated to the production of plastic parts – new procedures were conceived and a major attention was oriented to the quality standards, preventive maintenance planning and production forecasts. These practices led to a reduction of the rejection rate, which lowers the required input of virgin plastics and electricity to produce a unit amount of good parts.

4.2.2.5 Particularities on Business Case 2

For the business case 2 – the Automotive Plastic Parts, it was not possible to evaluate the full potential of MONSOON because of few problems related to data collection from newly introduced in-mold sensors: some failures actually impacted on data quality. In fact, these problems were associated to the physical installation of sensors in the mold tool and the teams are now implementing new changes, in order to solve the correct housing of the sensors.

During this learning process of the sensor behavior, the trend analysis of the machine behavior (through predictive monitorization) allowed us to identify the root cause of the malfunction of in-mold sensors and work directly on the issue.

After the implementation of the changes, we will be able to stablish a correlation between the injection process and mold behavior, honing the injection balance and achieve better plastic parts and improvement of the injection cycle. This cycle time improvement will affect in the gains of production time, increasing the production efficiency in 2 – 3 %.

4.2.2.6 Other

Another relevant outcome has been that the project itself has brought the opportunity to identify some optimization potentials in the involved industries, has promoted the interaction among different departments as well as with few technology providers. For instance, the introduction of predictive control methods triggered also an improvement of an existing system performing quality control. In the end, the project activities generated a positive impact on the whole ecosystem.

4.3 Replicability and scalability evaluation

In the context of MONSOON project, the replicability means the ability to reproduce the same outputs of the system for the same input data and conditions. The most critical is the replicability of the predictive functions build on the historical datasets. This requires that the data provenance is recorded across to the whole platform, i.e. for each data record, it is necessary to track data source, and all operations applied on the data. For the MONSOON platform, this is implemented mainly by the Apache Nifi, which is the main technology for data routing on the site or in the Data Lab platform. When the data are sent to the Data Lab, they are sent together with the data provenance metadata tracked on the site, so in the Data Lab it is possible to track data records back up to the source in the production environment on the site. Besides the data provenance, an important aspect of the replicability is the versioning of software components which are used for the data processing, since with the different version, it is possible to generate different outputs. The whole platform is deployed using the Docker technology and Python-stack packages, which ensure explicit control of the software versions.

By definition, the scalability is the property of the system to handle a growing amount of data by adding resources to the system. In the case of the MONSOON platform, the data can be characterized by Big Data properties such as the volume of data in bytes, the velocity of updates or a variety of heterogeneous data. For the Real-time plant operations platform deployed on the site, the most important scalability property is the velocity of data updates measured by number or data records processed per second and latency of the processing. The processing time includes the time needed to fetch the input data from the data source, pre-processing of data and routing them to the Runtime Container and Data Lab, computation of the prediction by the predictive functions and routing of the outputs back to the production environment. The most critical components for routing of data are implemented using the Apache Nifi technology, which can be deployed in the cluster and allows horizontal scaling of the Real-time plant operations platform by adding new servers into the cluster. According to the benchmark testing, Apache Nifi technology can handle data flow up to the 10 000 of data records per 1s with low latency, depending on the communication protocol used to access data from the sources. For the computing of predictions, the scalability is constrained mainly by the implementation of the predictive functions, which depends on the complexity of the implemented method. This can range from relatively simple logistic models or decisions trees up to the complex deep-learning models with thousands of parameters. For complex deep learning methods, the MONSOON adopts the Python frameworks such as Tensorflow which can be highly optimized by computing on the GPUs. The number of deployed functions is scaled-up by the Docker technology since the Runtime container is executing each predictive function in the separated Docker container.

For the Cross-sectorial Data Lab platform, the most important data property is the total volume of collected data from all sites. The Data Lab is usually deployed in the multi-tenant configuration where the data from one site can be stored and processed on the isolated cluster of servers, and each tenant can be scaled separately. The most critical is the Data Lab data storage characterized by read/write performance and the Data processing framework characterized by volume, which can be processed by various data analytical methods. The main technology used to implement data storage in the Data Lab is the Apache

Cassandra NoSQL database with the KairosDB frontend to process time-series data. The database can be deployed in the cluster and can be horizontally scaled. According to the existing production deployments, Apache Cassandra technology can handle hundreds of TBs with thousands of requests per 1s. The Data processing framework is based on the Python technological stack where the methods are implemented as the scripts. The scripts are executed in the separated containers and scaled by Docker technology. Additionally, for very large datasets (hundreds of TBs), it is possible to implement data processing using the distributed data processing framework such as Apache Spark deployed on the cluster of servers.

According to data collected during the project period for pilots, typical data flow in the runtime environment on the site can vary from 100 records per minute to 1000 records per 1s with the volume up to the hundreds of MBs per 1s. Data from the site are usually sent to the Data Lab with the 15 minutes latency, but this can be configured up to the 1s latency. The overall volume of the historical data collected for one site is about 400 GB per year, so generally, the Data Lab platform deployed on the moderately-sized cluster with up to the 20 data nodes can reliably handle hundreds of production sites.

5 Conclusion and further improvements

This deliverable presents the final impact, for dataflow simulation, but also for the defined KPIs for both Aluminium and Plastic domain.

The analysis of the KPIs calculation and its associated results reveals that the selected KPIs are aligned with the scope and the objectives are measurable, either to the industrial domains as to the proposed goals of the MONSOON project. However, along with the maturity of the developments, some improvements need to be considered to better characterize the scenario of each pilot.

It is certain that the MONSOON solution is scalable to other domains / segments and the target is to perform a standard tool to be, then, personalized for different applications. The MONSOON solution as a service is also business-oriented and deliver functionalities that the end-user can easily visualize business impacts and address them to his procedures, in order to identify the causes and establish actions to reduce those impacts.

Acronyms

Acronym	Explanation
KPI	Key Performance Indicators
LCA	Life Cycle Assessment
AP	Aluminum Pechiney
LCEN	Life Cycle Engineering
CAP	Capgemini

List of tables