processes

MDPI

Article

Impact of Data Grouping on the Multivariate Analysis of Several Concrete Plants

Malika Perluzzi [1], William Wilson [2] and Ryan Gosselin [1,*]

[1] Department of Chemical & Biotechnology Engineering, Faculty of Engineering, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada; malika.perluzzi@usherbrooke.ca
[2] Department of Civil and Building Engineering, Faculty of Engineering, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada; william.wilson@usherbrooke.ca
* Correspondence: ryan.gosselin@usherbrooke.ca

Abstract: Multivariate analysis can be used to study industrial process data exhibiting collinearity between variables. Such data can often be collected into conceptually meaningful groups or blocks. While data blocks may appear intuitive (e.g., raw material properties vs. process parameters), such blocking is sometimes much more subjective. The novelty of this work lies in the investigation of the impact of data blocking on the subsequent analysis. To our knowledge, no such investigation can be found in the literature. To fill this gap, we analyze the impact of grouping data from 10 Canadian concrete plants in which multiple blocking alternatives are considered. The analysis is performed via principal component analysis (PCA) to reduce the dimensionality of the matrix and also via consensus principal component analysis (CPCA). The data grouping options are as follows: (1) all data combined into a single block, (2) grouped according to the factory, (3) grouped according to parameter type, and (4) grouped according to parameter type within each factory. The results show that the grouping strategy alters the conclusion by emphasizing specific aspects of the data. While some grouping options emphasized seasonal variations, others emphasized other characteristics in the data, such as step changes in processing regimes or the significant impact of the raw materials' moisture on the process. As such, it appears relevant to consider multiple blocking options when analyzing complex datasets. Doing so will give the analyst a better understanding of overarching trends and more subtle characteristics of the dataset.

Keywords: multivariate analysis; PCA; CPCA; data grouping; concrete plant

Citation: Perluzzi, M.; Wilson, W.; Gosselin, R. Impact of Data Grouping on the Multivariate Analysis of Several Concrete Plants. *Processes* **2023**, *11*, 1551. https://doi.org/10.3390/pr11051551

Academic Editor: Jie Zhang

Received: 29 March 2023
Revised: 13 May 2023
Accepted: 15 May 2023
Published: 18 May 2023

1. Introduction

The Fourth Industrial Revolution, more commonly known as Industry 4.0, brought on the automation of production processes through the massive use of electronics, information, and communication technologies [1]. This automation is achieved with computers, sensors, actuators, and control processing units [1,2]. These devices are interconnected and can communicate with one another [2,3]. However, one of the most important aspects of the Fourth Industrial Revolution is certainly the rise of big data [2]. Big data can be defined as a large volume of complex and growing datasets obtained from multiple sources [2,4]. One of the ways to analyze such datasets is through latent variable multivariate analysis [5–8].

Multivariate analysis can be used to study a single dataset exhibiting collinearity or correlation between variables [9]. Industrial process data typically contain significant levels of correlation between variables, which explains why multivariate analysis is commonly used to perform process monitoring [9]. Principal component analysis (PCA) is a standard latent variable method commonly used to reduce the dimensionality of multivariate datasets and assist in visualization. It has become popular in industrial process monitoring as a standard and basic method [9]. In the case of process data, PCA has many applications including, for example, the detection of leaks from a boiler at Syncrude Canada's utility plant [10] and the study of various concrete mixes to determine the elements that can

influence the concrete properties [11]. Among other things, PCA allows the monitoring of the stability of an industrial process by comparing one or more process states with specified lower and upper control limits [12].

Furthermore, multivariate analysis is widely used for classification purposes. Statistical classification consists of assigning a class or category to a data or a data group. During data processing, methods such as PCA help render the relationships between variables more intelligible [13]. Wang et al. have also used several multivariate methods, including PCA, to categorize the chemical composition of five kinds of crop straw [13]. Zapata et al. conducted a similar study, but rather than comparing the same product type, they focused on studying different ones. Raman imaging and multivariate analysis have been used to discriminate textile fibres and fabrics [14].

Rather than comparing process states within a single dataset, multivariate analysis can also be used to compare diverse datasets. For example, Kruszewski and Obiedziński proposed a study analyzing the composition of raw materials used in the chocolate manufacturing process from three different producers [15]. Although these three manufacturers all produced chocolate, fundamental differences are raised at the manufacturing process technology level and among the process parameters [15].

In terms of applications of multivariate analysis, there seem to be three distinct scales. The first is to analyze a single process or dataset. The second aims to analyze multiple similar datasets, such as parallel process lines. Finally, the third scale is to compare multiple diverse datasets, such as differing processes, or multiple factories. However, producers often present significant differences, making them difficult to compare.

To illustrate our general ideas on process monitoring, we have chosen to illustrate them using datasets from concrete production. Compared to many industrial processes, the industrial concrete process has significant potential for improvement. Specifically, many efforts are devoted to concrete research, but most remain at the laboratory scale [16–18]. This is mainly attributable to the laws and standards that dictate concrete production [19]. Therefore, concrete production is homogeneous in the factory. Because of this relative homogeneity between some producers, datasets from different concrete producers may be compared to one another.

In this paper, we propose to analyze the impact of data grouping on the conclusions drawn from latent variable multivariate analysis. While data grouping may sometimes be straightforward, this is not always the case, and subjective choices may impact conclusions. The methods selected here are principal component analysis (PCA) and consensus principal component analysis (CPCA). While PCA considers relationships between individual variables, CPCA also considers relations between groups, or blocks, of variables [20]. While many methods could have been used in this work, PCA and CPCA were chosen in this work because of their widespread use in the field of chemometrics. We use data from multiple concrete producers to compare results and draw conclusions from each data grouping.

This paper first introduces experimental methods and presents the data as well as the manipulations and pre-processing carried out. A brief review of PCA and CPCA is then presented followed by the results obtained with the different data grouping options.

2. Materials and Methods

2.1. Raw Data Matrix

Data from 10 Canadian concrete plants were provided by Marcotte Systems. For each concrete plant, 11 variables were acquired. These data were found to pertain to three broad categories. The first category is related to solid compounds, which include the weight of cement (1), water (2), and admixtures (3) (i.e., admixtures are products added to the concrete mix that cause changes in the properties of the mixture in the fresh or hardened state [21]) and aggregates. Aggregates are subdivided into two types: fine (4) and coarse (5). The second category is moisture, which includes the moisture of both types of aggregates: fine (6) and coarse (7). The third category is the target recipe, which represents setpoints

fixed by the concrete producer. This category includes the air content (8), the slump (9), the mix strength (10), and the volume per batch (11). In total, the dataset contains the 11 variables distributed among the three categories over a four-year period (2015–2019). This represents a data matrix for each of the 10 concrete plants of X (2711 × 11) containing 2711 time-points, representing a daily data acquisition rate. While the 10 plants are operated and run by two companies, the data matrix of each plant is comparable and encompasses the same 11 variables over the same time-period. As such, they all have the same dimensions (2711 × 11). Recall that the data presented are time series.

2.2. Data Grouping

After obtaining the data, the analyst must decide how to organize the information. In this work, we propose to investigate the impact of data blocking on PCA and CPCA. As such, data grouping was required. There are many ways to group data, four of which were selected in this work, as presented in Figure 1.

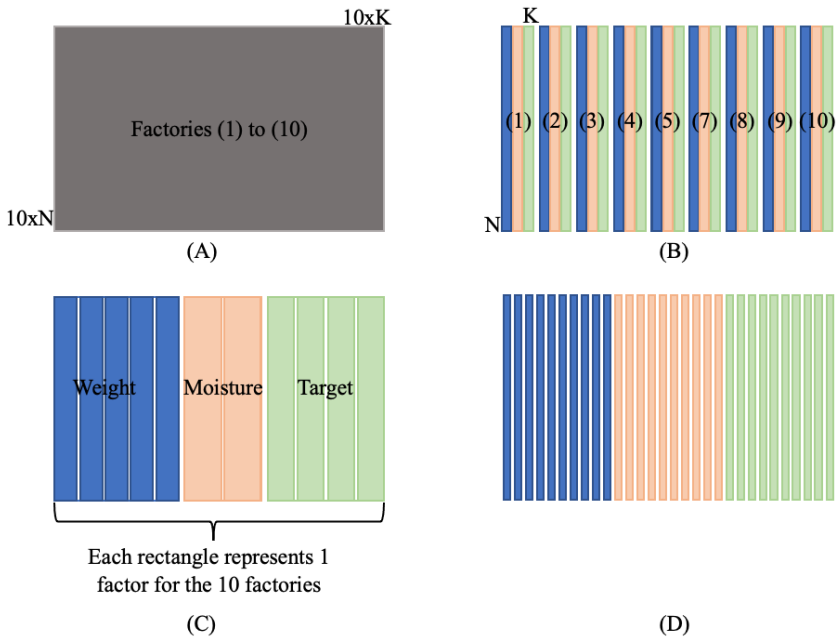


Figure 1. Data grouping options: (**A**) all factories are combined, (**B**) factory grouping, (**C**) parameter grouping, and (**D**) splitting factories and parameters.

The first option shown in Figure 1A combines the ten matrices (2711 × 11), one from each concrete plant, to obtain a single large matrix (2711 × 110). The second option (Figure 1B) subdivides the data according to the ten concrete plants regardless of the data type (10 blocks of 2711 × 11). The third data grouping option (Figure 1C) subdivides the data according to the three types of data regardless to the concrete plant: weight, moisture, and target (three blocks with sizes of 2711 × 50, 2711 × 20, and 2711 × 40). The final option (Figure 1D) subdivides the data based on both concrete plant and data type. This option results in 30 blocks, each with dimensions ranging between 2711 × 2 and 2711 × 5.

Following this subdivision of the data, the first option will be analyzed via principal component analysis (PCA), whereas the three other options will be analyzed via consensus principal component analysis (CPCA).

2.3. Data Pre-Treatment

All available data underwent pre-treatment before performing multivariate analyses. First, the data were centred and scaled to obtain a mean of zero and a standard deviation of one. The data must undergo such pre-treatment due to the different ranges of measurement

units and numerical values present in the data. This pre-treatment was performed on all the columns of the matrix by subtracting the mean and dividing by the standard deviation [22]:

$$X_{CS} = \frac{x_1 - \bar{x_1}}{std(x_1)}, \ldots, \frac{x_k - \bar{x_k}}{std(x_k)} \quad (1)$$

where k represents each of the columns of the data matrix, while x represents all observations in each column. Subsequently, the respective weights of each of the blocks were standardized (i.e., scaled). This reduction aims to balance the respective contribution of each block, sometimes having different sizes [23].

$$X_i = \frac{X_i}{\sqrt{k_i}} \quad (2)$$

In the equation above, the index i represents each of the data blocks. Finally, the treatment of outliers is required since their presence can bias the model. Outliers were removed from the dataset with Hotelling T^2 and Square Prediction Error (SPE) distance [24]. In particular, the joint use of Hotelling T^2 and SPE in a graph made it possible to visualize the location of the problematic points according to all the principal components. The problematic points are those with a large error (high SPE) or those whose projection falls far from the plane's centre (high T^2).

3. Overview of PCA and CPCA

3.1. Principal Component Analysis (PCA)

Principal component analysis corresponds to the basis of multivariate data analysis. This method is often used to observe tendencies and variations in a dataset containing correlated variables. This method decomposes a multidimensional dataset into a space with fewer dimensions [25]. This decomposition is obtained by using linear combinations of variables from the original data set to represent them as new variables [25,26]. In practice, PCA represents the data matrix as a product of two matrices [25].

$$\mathbf{X} = \mathbf{T} \times \mathbf{P^T} + \mathbf{E} \quad (3)$$

The data matrix **X** (M × K) containing M observations and K variables is decomposed into a loadings matrix **P** (K × A) and a scores matrix **T** (M × A), where A is the number of principal components. The rows of the loadings matrix represent linear combinations of the variables of the **X** matrix and express the relationships between the variables [26]. The columns of the scores matrix represent the observations based on the new system of variables induced by the loadings [26]. Finally, the matrix **E** (M × K) is the error matrix representing the prediction error with respect to **X** [26]. Note that this matrix has the same dimensions as the initial matrix **X**. Normally, the analysis is carried out with data from the **X** matrix previously centred and scaled. A simplified diagram of the decomposition within the PCA method is represented in Figure 2 [20].

PCA can also be used to identify outliers within a dataset [27]. To do so, the distance of each observation from the centre of the dataset can be computed as a weighted Euclidean distance according to the variance of each principal component (Hotelling T^2) [27–30]:

$$T^2 = \mathrm{diag}\left(\mathrm{Tinv}(S)T^T\right) \quad (4)$$

$$S = \frac{T^T T}{N} \quad (5)$$

where A corresponds to the number of principal components, N to the number of observations and S is associated with the variance. Then, the orthogonal projection distance based on an error point of the PCA model is calculated as follows [24]:

$$SPE = \sum E^2 \quad (6)$$

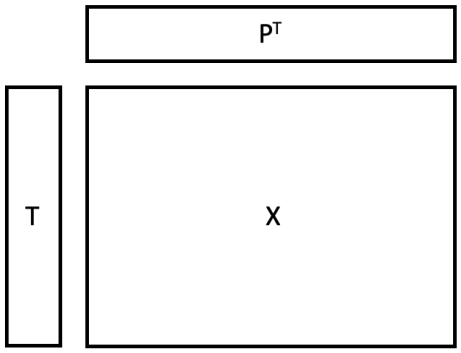


Figure 2. Decomposition of the matrix **X** into a product of the scores (**T**) and the transpose of the loadings ($\mathbf{P^T}$) with the PCA method.

3.2. Consensus Principal Component Analysis (CPCA)

CPCA is an extension of PCA used to compare several blocks of variables measured on similar objects, improving the interpretation of multivariate models. These blocks are created by the user based on the availability of additional information, allowing the variables to be grouped into conceptually significant groups [20].

The data are first divided into B blocks ($\mathbf{X_1} \ldots \mathbf{X_B}$). Then, the data must undergo pre-processing, including centring and scaling them, as presented in Equation (1). Block scaling is also required to balance their respective contribution to the model since the blocks can be of different sizes, as presented in Equation (2). The matrix is then represented as scores and loadings, like in PCA. However, in CPCA, there are two levels of scores: scores and super-scores. Scores represent relationships between observations based on the variables within each block, whereas super-scores provide an overall representation of the observations using the full dataset (Figure 3). Each block contains loadings (**P**) as in PCA. The relative importance of each block in the projection of the super-scores is represented by the block weights (**W**) [20].

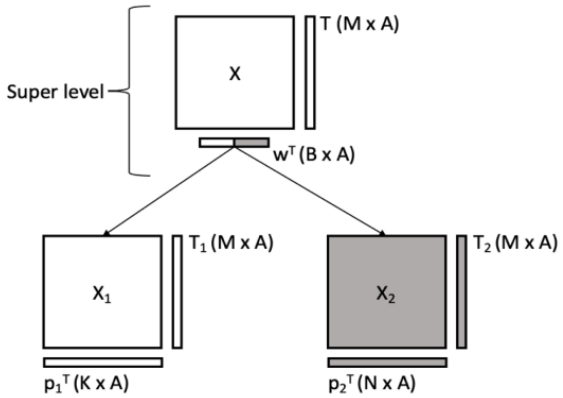


Figure 3. Matrix decomposition illustrated for two blocks with CPCA method.

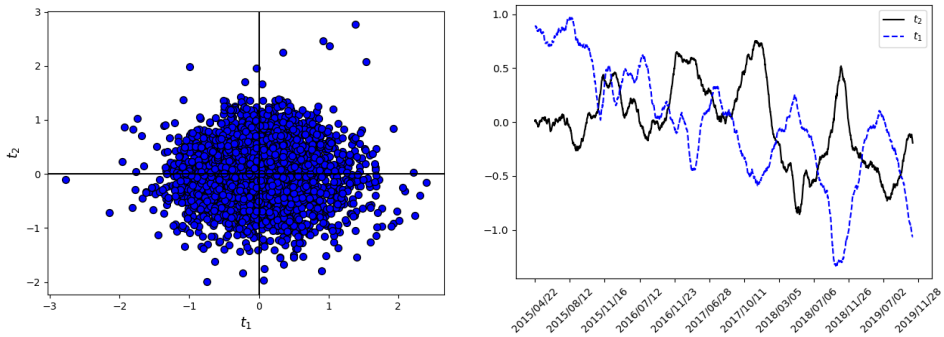
In the figure, A corresponds to the number of principal components, B to the number of blocks, M to the number of observations in each block, and N and K to the number of variables in each block, respectively.

4. Results and Discussion

4.1. Data Grouping Option 1—Combined Factories

Principal component analysis is often used to observe trends or variations within a dataset. This method was therefore applied to the first data grouping option, which combines data from all the factories into a single large matrix (Section 2.2). The figure below presents the results of the PCA scores.

Only the two first PCA principal components ($\mathbf{t_1}$ (4.1%) and $\mathbf{t_2}$ (3.4%)) are presented in Figure 4, as they present the greatest sources of variance. These low values are caused by the great complexity of the dataset and the fact that it comes from 10 independent manufacturing sites. While all sites are independent, we have chosen to combine them as they share similar realities (e.g., all require similar raw materials, are subject to similar market pressures, and face similar seasonal variations). Recall that scores represent observations within the dataset. By plotting $\mathbf{t_1}$ vs. $\mathbf{t_2}$ (Figure 4A), it is possible to visualize possible relationships between observations. An observation located near the centre (0.0) corresponds to a typical observation, while one located far from the centre represents an atypical observation. Furthermore, observations located in the same region of the score plot show similarities, whereas observations located in opposite regions of the score plot show differences. Figure 4A illustrates a large clump of data clustered around the centre in which it is relatively difficult to identify trends and draw conclusions.



(**A**) (**B**)

Figure 4. Scores plots shown in the representations of the two first principal component (**A**) and the two first principal component in function of time (**B**).

For this reason, the scores are also presented as a function of time in Figure 4B. Note that data presented in this figure are smoothed using the Savitzky–Golay algorithm to improve legibility. The first two principal components show notable temporal variations that are consistent with seasonal changes. Such seasonality is not surprising, as temperature is known to directly influence the concrete batching process (e.g., the humidity of aggregates, the adjustment of water, etc.). As mentioned, all plants are located in Canada and are expected to experience similar temperature fluctuations.

The loadings plot makes it possible to visualize relationships between variables. On the one hand, a variable located near the centre is of limited use for the model as it lacks the ability to discriminate between observations within the dataset. On the other hand, a variable located far from the centre is very useful since it has great discriminating power. Furthermore, variables located near one another are positively correlated and are likely to have similar effects on the observations, while variables located in opposite regions of the plot are negatively correlated and have opposite effects.

Figure 5 presents the combined loadings ($\mathbf{p_1}$ (4.1%) and $\mathbf{p_2}$ (3.4%)) of all variables in all concrete plants. As mentioned, all the plants have the same variables: the weight of cement, the weight of water, the weight of admixtures (Admix), the weight of fine (Sand) and coarse aggregates (Agg), the moisture of fine (Moisture-S) and coarse aggregates (Moisture-Ag), the air content, the slump, the mix strength (MS), and, finally, the volume

per batch (Vol). Figure 5 can be used to determine if any concrete plant or any variable may explain trends in the data. However, the number of variables present in this graph complicates the interpretation of general trends. For example, one can clearly see that the weight of sand in factory B4 (variable Sand_B4, located at $p_1 = -0.27$) stands out, but it is much less obvious to determine if the weight of sand in general (for all 10 factories) or if the factory B4 (as a whole) stands out. Therefore, the Hotelling T^2 metric presented Table 1 is used for a better visualisation of the information.

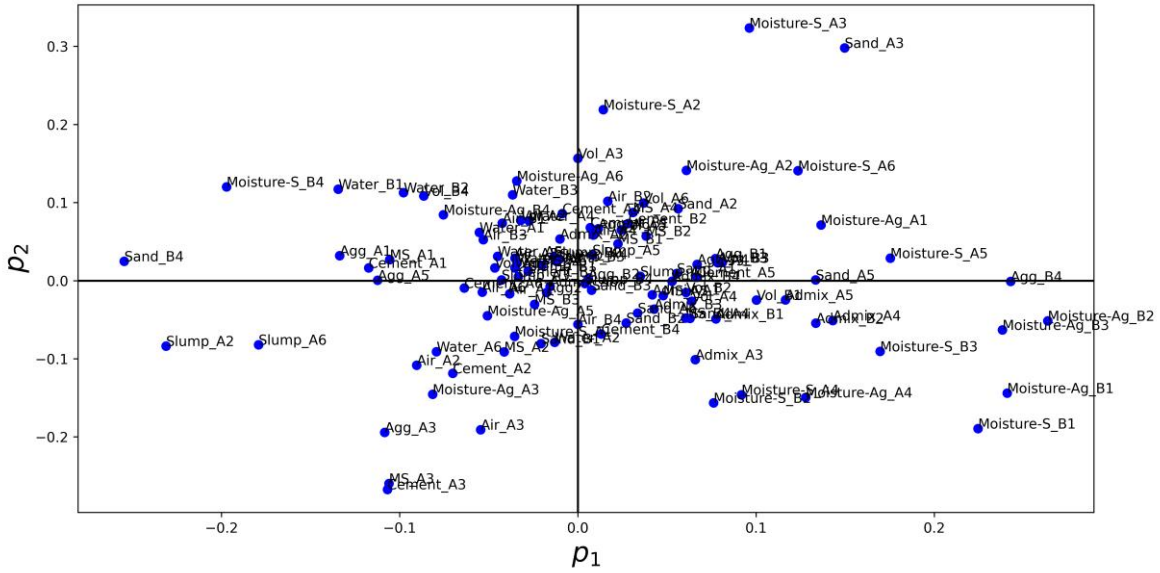


Figure 5. Loadings plot of all variables in all concrete plants.

Table 1. Hotelling T^2 metric for each concrete plant and each variable.

	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	Sum for Each Variable
Admixture (Admix)	0.23	0.32	1.60	2.54	1.56	0.03	0.92	2.29	0.35	0.31	10.14
Cement	1.55	2.09	9.11	0.82	0.49	0.46	0.51	0.68	0.13	0.53	16.36
Water	0.76	0.70	0.18	0.73	0.33	1.60	3.50	2.45	1.48	0.17	11.89
Sand	0.35	1.28	12.22	0.69	1.96	0.31	0.76	0.40	0.02	7.19	25.20
Sand moisture (Moisture-S)	0.70	5.29	12.53	3.28	3.47	3.86	9.49	3.33	4.06	5.86	51.86
Aggregate (Agg)	2.08	0.06	5.44	0.54	1.39	0.50	0.74	0.001	0.73	6.49	17.98
Aggregate moisture (Moisture-Ag)	2.61	2.60	3.05	4.25	0.51	1.92	8.67	7.94	6.68	1.41	39.64
Volume (Vol)	0.27	0.77	2.70	0.53	0.53	1.24	1.17	0.43	0.78	2.12	10.53
Air	0.19	2.19	4.33	0.38	0.23	0.34	0.79	1.17	0.61	0.34	10.57
Slump	0.14	6.65	0.20	0.10	0.13	4.27	0.10	0.12	0.08	0.001	11.79
Mix strength (MS)	1.31	1.10	8.66	0.94	0.29	0.09	0.30	0.52	0.17	0.65	14.03
Sum for each plant	10.16	23.06	60.03	14.80	10.88	14.62	26.97	19.33	15.09	25.07	Total

Hotelling T^2 is used to quantify the distance from the centre and can be used to distinguish which variables stand out. These values were summed by variable and by concrete plant to produce an overall value. The idea is that the higher the sum per variable or per concrete plant, the higher the discriminating power.

Table 1 shows that the factories with the highest Hotelling T^2 values are A2, A3, B1, and B4. These plants stand out from their counterparts because of their significantly higher discriminatory power, implying that the processing conditions inside these plants varies significantly over time. On the other hand, the variables with the highest Hotelling T^2

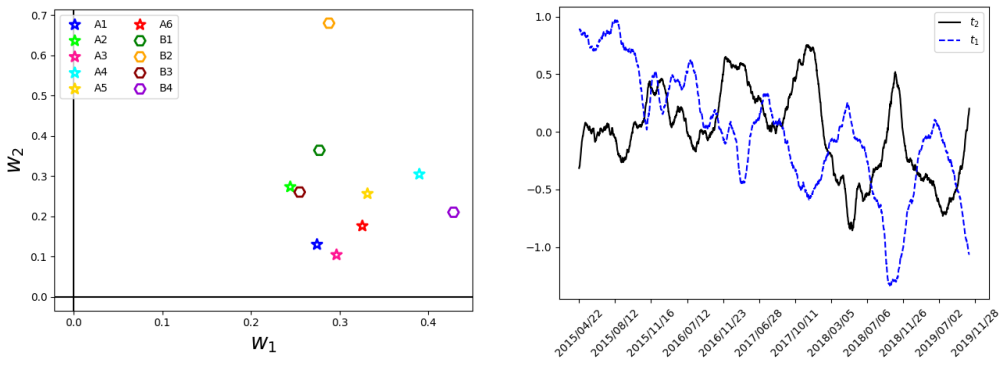
value are the quantity of sand in the mixture and the moisture of the sand and coarse aggregates. The amount of water in a concrete mix is a crucial aspect since this modifies the water/cement ratio, directly impacting the concrete's quality [31]. However, moisture has greater discriminating power than water. This also means that the amount of water does not vary significantly over the time interval. In other words, the moisture determines at what time of the year the different concrete recipes take place. The sand and aggregate moisture have a high Hotelling T^2 value for all factories. This is not the case for the quantity of sand, which is high only for the A3 and B4 factories.

4.2. Data Grouping Option 2—Factory Grouping

CPCA is often used to compare several blocks of descriptive variables measured on similar objects, improving the interpretation of multivariate models [20]. This method is applied to the second data grouping option, which consists of creating 10 blocks, each representing a concrete plant regardless of data type.

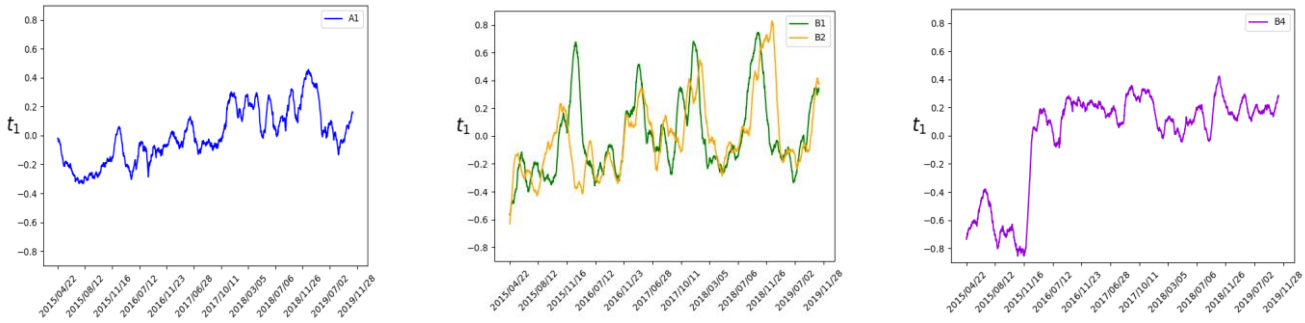
Recall that the block weights represent the relative importance of the different blocks in the model. Figure 6A shows that some plants are positioned closer to the center (0,0) while others are located along the vertical or horizontal axis or a mixture of both. For example, factory A1 has the shortest distance from the centre. Then, factory B2 can be explained almost exclusively with $\mathbf{w_2}$ (3.4%). On the contrary, factory B4 is explained almost exclusively with $\mathbf{w_1}$ (4.1%) only. Figure 6B presents the overall temporal variations experienced by all factories ($\mathbf{t_1}$ and $\mathbf{t_2}$ scores for the full dataset). The interesting aspect is that the trends are the same as those observed for the first data grouping option in Figure 4B. However, the global time trends presented are insufficient to entirely explain seasonal phenomena. To be able to explain seasonality more adequately, three distinctive scenarios are presented in Figure 7. The first case presents a factory located near the centre of the $\mathbf{w_2}$ vs. $\mathbf{w_1}$ plot. The second case presents factories located at high $\mathbf{w_2}$ values. The third case is a factory located at a high $\mathbf{w_1}$ value.

Firstly, temporal trends in plant A1 (Figure 7A) do not appear to present any structured variations. Rather, the data of the time seem relatively noisy. These findings are also consistent with the positioning of concrete plant A1 in Figure 6A. This plant is relatively close to the centre (0.0), indicating its low discriminating power. Secondly, factories B1 and B2 illustrated in Figure 7B have several seasonal oscillations over time. Nevertheless, these oscillations show a certain phase shift, indicating that these oscillations are not purely seasonal and relate to other factors, as yet unknown. Moreover, in Figure 6A, these factories are positioned at the extreme vertical ($\mathbf{w_2}$), indicating that they have greater discriminating power. Thirdly, concrete plant B4, shown in Figure 7C, includes an interesting step change in its time trend. Like plants B1 and B2, plant B4 is located away from the center in Figure 6A, indicating more significant discriminating power.



(**A**) (**B**)

Figure 6. Block weights of the factories grouping (**A**) and the two first principal component in function of time (**B**).



(**A**) (**B**) (**C**)

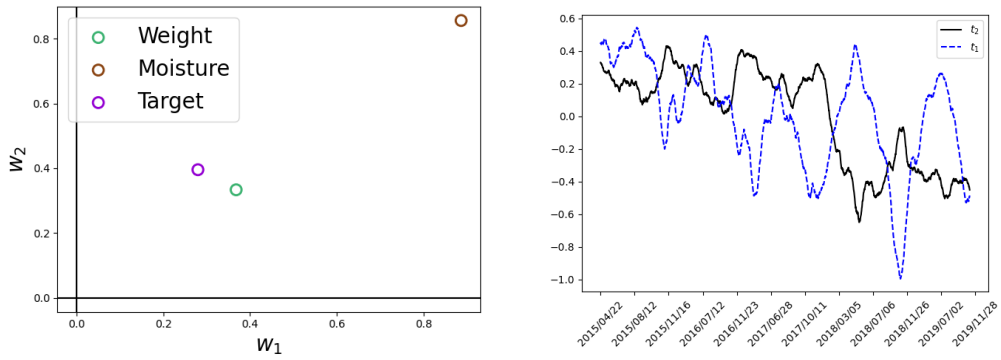
Figure 7. First principal component (t_1; 3,1%) of factories A1 (**A**), B1 and B2 (**B**) and B4 (**C**) factories as a function of time.

In short, the positioning of $\mathbf{w_1}$ and $\mathbf{w_2}$ of the plants relative to the centre in Figure 6A make it possible to criticize their discriminating power and the structured temporal structure they possess. Although positioning relative to $\mathbf{w_1}$ and $\mathbf{w_2}$ confers greater discriminating power than plants near the centre, these two components ($\mathbf{w_1}$ and $\mathbf{w_2}$) explain changes of a different nature. In the current dataset, $\mathbf{w_2}$ is associated with seasonal changes, while $\mathbf{w_1}$ is associated with significant regime shifts. These conclusions differ from those previously drawn using the Hotelling T^2 metric for each concrete plant and each variable (Table 1). Therefore, the method selected to group the data does not allow similar conclusions about the factories or the most discriminating variables.

4.3. Data Grouping Option 3—Parameter Grouping

CPCA analysis is applied to the third option, which consists of grouping the data according to the types of parameters.

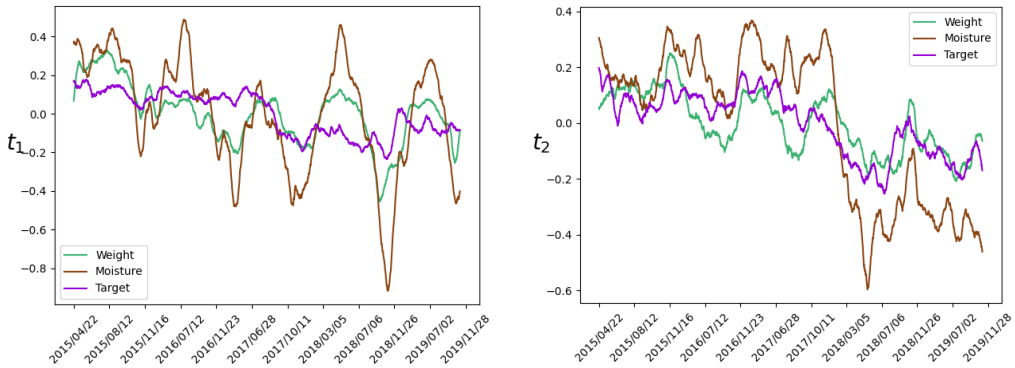
Figure 8A shows that the blocks associated with the weight and the target are very close. On the contrary, the block associated with moisture is found entirely in the far corner of the plot. Therefore, its position in the graph indicates that this block has great discriminatory power. In other words, moisture has the greatest discriminating power between observations. This also partially corresponds to the conclusions drawn in Table 1 from the first data grouping option. In this sense, in both cases, the moisture of fine and coarse aggregates has great discriminating power. This is mainly due to the seasonal variations that the factories undergo. However, this option does not make it possible to distinguish factories or even specific atypical parameters.



(**A**) (**B**)

Figure 8. Block weights of the parameters grouping ($\mathbf{w_1}$; 5.5%, $\mathbf{w_2}$; 4.5%) (**A**) and the two first principal component in function of time (**B**).

The seasonal variations are also noticeable in Figure 8B. However, global time trends are not sufficient to explain seasonal phenomena. The first two principal components ($\mathbf{t_1}$ (2.2%) and $\mathbf{t_2}$ (1.5%)) for each block were plotted as a function of time in Figure 9.



(**A**) (**B**)

Figure 9. First two principal components ($\mathbf{t_1}$ (2,2%) (**A**) and $\mathbf{t_2}$ (1,5%) (**B**)) of each block as a function of time.

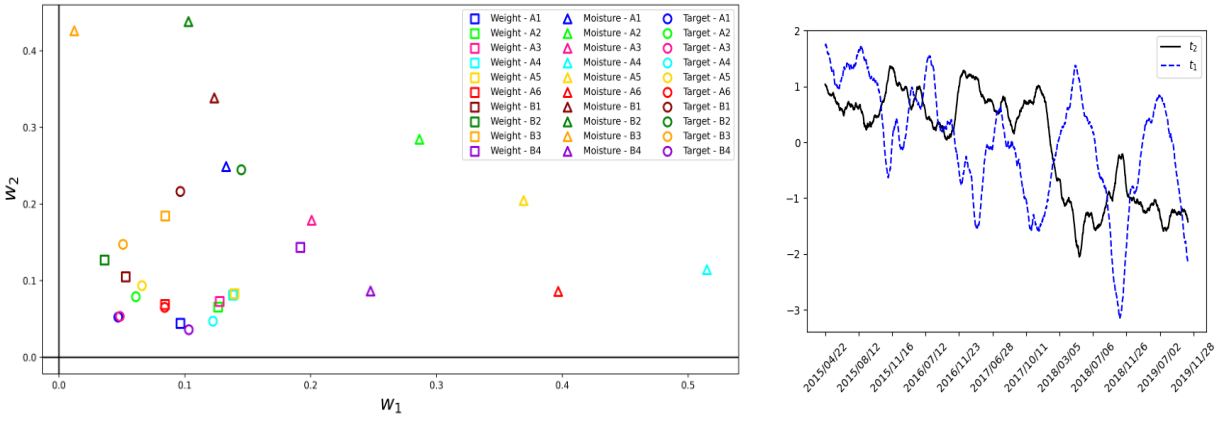
The figure above shows the temporal variations in each of the three data blocks. As the CPCA was computed with two components, Figure 9A illustrates the trends associated with the largest source of variance, whereas Figure 9B does so for the second largest. Figure 9 illustrate that the seasonal changes observed in Figure 8B are essentially induced by moisture. Figure 8A has led to the conclusion that moisture has the most significant source of variance, which is confirmed by Figure 9. Therefore, moisture explains the perceived step in the second data grouping's temporal variation (Figure 6B). However, the impact of moisture is not obvious at first glance since the moisture measurements are distributed within each block. In the opposite case, i.e., considering moisture as a block (third data grouping) in its own right makes it possible to visualize its impact on temporal variations.

4.4. Data Grouping Option 4—Splitting Factories and Parameters

CPCA analysis is applied to the fourth data grouping, which splits the data according to the factories and parameters.

Figure 10A show 30 blocks because the data have been split according to the concrete plant and their data type. First, most of the blocks connected to the weight and the target are close to each other, which indicates a correlation between them. In addition, they are positioned relatively close to (0.0), indicating a low discriminatory power between observations. These conclusions coincide with those drawn using the third grouping of data. However, the blocks associated with the weights of the factories raised at data grouping option two (B1, B2, and B4) are found further from the centre, indicating their greater discriminating power. In addition, using the conclusions drawn from data grouping option one, it is possible to guess that these blocks of masses are distinguished due to the sand dosage.

On the other hand, moisture blocks are found at the end of the graph. Their location indicates that humidity has a significant discriminating power. This conclusion is the same as that drawn for the first and third data grouping. Here again, this is mainly due to the seasonal variations that the factories undergo. These seasonal variations are also noticeable in Figure 8B. Incidentally, the time trends presented in this graph are the same as those in the third data grouping option.



(**A**) (**B**)

Figure 10. Block weights of splitting factories and parameters ($\mathbf{w_1}$; 5.5%, $\mathbf{w_2}$; 4.5%) (**A**) and the two first principal component in function of time (**B**).

5. Conclusions

Data can often be presented in meaningful groups of variables, often called blocks. The novelty in this work is investigating the impact of data blocking on PCA and CPCA. As such, data grouping was required. To do so, datasets from 10 Canadian concrete plants are used. The data pertain to three broad categories: weight, moisture, and production targets. While data grouping may sometimes be straightforward, this is not always the case. In this case, four grouping alternatives were selected in this work. The first option combines the ten matrices, one from each concrete plant, to create a single large matrix. The second option subdivides the data according to the ten concrete plants regardless of the data type. The third data grouping option subdivides the data according to the three data types, regardless of the concrete plant. The final option subdivides the data based on the concrete plant and data type.

The results of the first data grouping made it possible to observe seasonal variations. Although these variations are difficult to interpret, it was possible to identify specific plants and variables as being of interest. The results of the second data grouping made it possible to represent the dynamic behaviour of the plants better. These results also made it possible to better understand the seasonal variations resulting from the first data grouping. The major advantage of this data grouping lies in observing the temporal evolution of each factory. The third data grouping shows the significant impact of moisture. Moreover, this method allowed us to observe a step change in the data, which is only visible now. Finally, the fourth data grouping provides conclusions similar to the third. However, this method makes it possible to visualize the great moisture variability for all plants.

Overall, this work illustrates the importance of data processing on the analysis and the interpretation. This type of analysis can also be used to improve the process under study. Such points are broadly applicable to any sufficiently large dataset, be it industrial in nature or not.

Author Contributions: Conceptualization, M.P., W.W. and R.G.; methodology, M.P., W.W. and R.G.; software, M.P.; validation, M.P. and R.G.; formal analysis, M.P.; investigation, M.P.; data curation, M.P.; writing—original draft preparation, M.P.; writing—review and editing, M.P., R.G. and W.W.; visualization, M.P.; supervision, R.G. and W.W.; project administration, R.G. and W.W.; funding acquisition, R.G. and W.W. All authors have read and agreed to the published version of the manuscript.

Funding: Funding was received from Mitacs and Marcotte Systems (#IT23720).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are unavailable due to privacy.

Acknowledgments: We would like the acknowledge the financial support from Mitacs and Marcotte Systems.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barreto, L.; Amaral, A.; Pereira, T. Industry 4.0 implications in logistics: An overview. *Procedia Manuf.* **2017**, *13*, 1245–1252. [CrossRef]
2. Karatas, M.; Eriskin, L.; Deveci, M.; Pamucar, D.; Garg, H. Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives. *Expert Syst. Appl.* **2022**, *200*, 116912. [CrossRef]
3. Ghobakhloo, M. Industry 4.0, digitization, and opportunities for sustainability. *J. Clean. Prod.* **2020**, *252*, 119869. [CrossRef]
4. Vaidya, S.; Ambad, P.; Bhosle, S. Industry 4.0—A Glimpse. *Procedia Manuf.* **2018**, *20*, 233–238. [CrossRef]
5. Lawrence, N. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2003; Volume 16. Available online: https://proceedings.neurips.cc/paper/2003/hash/9657c1fffd38824e5ab0472e022e577e-Abstract.html (accessed on 17 October 2022).
6. Ferrer, A.; Aguado, D.; Vidal-Puig, S.; Prats, J.M.; Zarzo, M. PLS: A versatile tool for industrial process improvement and optimization. *Appl. Stoch. Model. Bus. Ind.* **2008**, *24*, 551–567. [CrossRef]
7. Mehmood, T.; Ahmed, B. The diversity in the applications of partial least squares: An overview. *J. Chemom.* **2016**, *30*, 4–17. [CrossRef]
8. Clementi, S.; Cruciani, G.; Curti, G. Some applications of the partial least-squares method. *Anal. Chim. Acta* **1986**, *191*, 149–160. [CrossRef]
9. Li, Z.; Tian, L.; Yan, X. An ensemble framework based on multivariate statistical analysis for process monitoring. *Expert Syst. Appl.* **2022**, *205*, 117732. [CrossRef]
10. Sun, X.; Marquez, H.J.; Chen, T.; Riaz, M. An improved PCA method with application to boiler leak detection. *ISA Trans.* **2005**, *44*, 379–397. [CrossRef]
11. Kobaka, J. Principal Component Analysis as a Statistical Tool for Concrete Mix Design. *Materials* **2021**, *14*, 2668. [CrossRef]
12. Kazmer, D.O.; Westerdale, S.; Hazen, D. A Comparison of Statistical Process Control (SPC) and On-Line Multivariate Analyses (MVA) for Injection Molding. *Int. Polym. Process.* **2008**, *23*, 447–458. [CrossRef]
13. Wang, X.; Yang, Z.; Liu, X.; Huang, G.; Xiao, W.; Han, L. The composition characteristics of different crop straw types and their multivariate analysis and comparison. *Waste Manag.* **2020**, *110*, 87–97. [CrossRef] [PubMed]
14. Zapata, F.; Ortega-Ojeda, F.E.; García-Ruiz, C. Forensic examination of textile fibres using Raman imaging and multivariate analysis. *Spectrochim. Acta Part A: Mol. Biomol. Spectrosc.* **2022**, *268*, 120695. [CrossRef] [PubMed]
15. Kruszewski, B.; Obiedziński, M.W. Multivariate analysis of essential elements in raw cocoa and processed chocolate mass materials from three different manufacturers. *LWT* **2018**, *98*, 113–123. [CrossRef]
16. Zhang, C.; Hu, M.; Dong, L.; Gebremariam, A.; Miranda-Xicotencatl, B.; Di Maio, F.; Tukker, A. Eco-efficiency assessment of technological innovations in high-grade concrete recycling. *Resour. Conserv. Recycl.* **2019**, *149*, 649–663. [CrossRef]
17. Le Béton: Un Matériau Technologique Faisant L'objet D'une Intense R&D: Techniques de l'Ingénieur. Available online: https://www.techniques-ingenieur.fr/actualite/articles/le-beton-un-materiau-technologique-faisant-lobjet-dune-intense-rd-105038/ (accessed on 16 September 2022).
18. Ngo, H.T.; Kaci, A.; Kadri, E.H.; Ngo, T.T.; Trudel, A.; Lecrux, S. Energy consumption reduction in concrete mixing process by optimizing mixing time. *Energy Procedia* **2017**, *139*, 810–816. [CrossRef]
19. *CSA A23.1-14/A23.2-14*; Conseil Canadien des Normes. Standards Council of Canada: Ottawa, ON, Canada, 2015. Available online: https://www.scc.ca/fr/standardsdb/standards/27899 (accessed on 19 September 2022).
20. Westerhuis, J.A.; Kourti, T.; MacGregor, J.F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **1998**, *12*, 301–321. [CrossRef]
21. Adjuvants. Infociments. Available online: https://www.infociments.fr/betons/adjuvants (accessed on 22 September 2022).
22. Bro, R.; Smilde, A.K. Centering and scaling in component analysis. *J. Chemom.* **2003**, *17*, 16–33. [CrossRef]
23. Hassani, S.; Hanafi, M.; Qannari, E.M.; Kohler, A. Deflation strategies for multi-block principal component analysis revisited. *Chemom. Intell. Lab. Syst.* **2013**, *120*, 154–168. [CrossRef]
24. Nelson, P.R.C.; MacGregor, J.F.; Taylor, P.A. The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 1–12. [CrossRef]
25. Bro, R.; Smilde, A.K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831. [CrossRef]
26. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
27. Shabbak, A.; Midi, H. An Improvement of the Hotelling T 2 Statistic in Monitoring Multivariate Quality Characteristics. *Math. Probl. Eng.* **2012**, *2012*, 531864. [CrossRef]
28. Turner, T.F.; Collyer, M.L.; Krabbenhoft, T.J. A general hypothesis-testing framework for stable isotope ratios in ecological studies. *Ecology* **2010**, *91*, 2227–2233. [CrossRef]

29. Zhou, Z.; Wen, C.; Yang, C. Fault Detection Using Random Projections and k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. *IEEE Trans. Semicond. Manuf.* **2015**, *28*, 70–79. [CrossRef]
30. Cogdill, R.P.; Anderson, C.A.; Drennen, J.K. Process analytical technology case study, Part III: Calibration monitoring and transfer. *AAPS PharmSciTech* **2005**, *6*, E284–E297. [CrossRef]
31. Kosmatka, S.H. Association canadienne du ciment Portland. In *Dosage et Contrôle des Mélanges de Béton: Manuel D'applications, Méthodes et Matériaux*, 8th ed.; Association Canadienne du Ciment: Ottawa, ON, Canada, 2011.