Data Drift for Automatic FAIR-compliant Dataset Versioning in Large Repositories

Alba González–Cebrián*, Iulian Ciolacu†, Michael Bradford*,
Ciprian Dobre†, Horacio González–Vélez*
*Cloud Competency Centre, National College of Ireland, Dublin, Ireland.
†Faculty of Automatic Control and Computers, University POLITEHNICA of Bucharest, Bucharest, Romania.
Email: {alba.gonzalez-cebrian, michael.bradford, horacio}@ncirl.ie,
{iulian.ciolacu, ciprian.dobre}@upb.ro

***Abstract*—Construed as a shift in the distribution or structure of data over time, data drift can adversely affect the performance of machine learning models and data-driven decisions. This study examines two data drift metrics, denoted as $d_{E,PCA}$ and $d_{E,AE}$, that are derived from unsupervised ML models: the reconstruction error-based metrics of Principal Component Analysis (PCA) and Autoencoders (AE). To investigate the robustness of these metrics, we have systematically accessed time-series datasets from the European Data Portal. Our experiments have examined data versioning through three basic events: creation, update, and deletion. The results are summarised and aggregated for all datasets, and unsupervised analysis based on Robust PCA and AE has been performed to examine patterns within the impact of dataset characteristics on data drift detection and computational efficiency. Our results indicate that both metrics aligned closely in performance with new records, suggesting consistent drift detection under normal conditions with FAIR compliance. However, high-dimensional datasets posed challenges for both PCA and AE models. Update events revealed discrepancies between the two metrics, suggesting that non-linear shifts affected AE-based metrics more than PCA-based ones. Deletion events demonstrated the resilience of these metrics against data loss, but also revealed variability in the reliability of the PCA model; i.e., data drift metrics derived from PCA and AE can be effective but sensitive to certain dataset characteristics.**

***Index Terms*—Data drift, Machine Learning, FAIR principles, Principal Component Analysis, Autoencoders, Data versioning, Time series, Dataset Versioning**

I. Introduction

Data-driven decision-making has become central to modern business and research practices. Machine learning (ML) models, artificial intelligence (AI), and data analytics all rely heavily on large datasets that continuously evolve due to various factors such as changing user behaviours, seasonal trends, or even errors in data collection. This evolution can lead to what is known as *Data Drift* (a.k.a. dataset shift): a change in the distribution or structure of a dataset that can significantly impact the performance and reliability of ML models [1].

It is widely acknowledged that versioning of datasets is a complex activity as it involves detecting not only intrinsic divergences, including noise, variance, and dynamic changes, but also dissimilarities due to data collected and distributed in different encodings or mirrored onto different platforms [2]. However, empirical reproducibility relies on enabling researchers access to the exact version of a given data set to corroborate results.

For this reason, detecting and managing data drift is crucial for maintaining a detailed data versioning system that tracks information changes to improve the accuracy and relevance of the predictive models involved. Arguably, ignoring data drift can arguably lead to decreased model performance, misinterpretation of results, and flawed business or research decisions. Hence, considering the high dynamic of the current data versioning system, accurately tracking changes in data drift might be not only interesting but also critical in the near future. However, while various methods have been developed to detect data drift, there is no one-size-fits-all solution.

This paper builds on our recent work, in which we proposed three metrics to quantify data drift [3]. Although our initial work focusses on the FAIR principles [4], this continuation study aims to further explore the limitations and characteristics of these data drift metrics by considering the impact of the size of the data set and other factors while maintaining FAIR compliance. Exploring different types of data drift events, including creation, update, and deletion, we seek to gain a deeper understanding of how these metrics behave and determine how robust these metrics are under different scenarios and what adjustments might be needed to improve their reliability. Through these experiments, we will discuss i) how dataset dimensionality affects the sensitivity and reliability of data drift metrics; ii) the influence of different data structures on the metrics' performance, particularly in cases with data scarcity; and, iii) the differences in metrics' computational efficiency.

The results of this study will not only provide a deeper understanding of data drift and suggest potential pathways for developing more robust data drift detection methods. Ultimately, our goal is to contribute to a more standardised approach to data versioning and drift detection, which can benefit data scientists, researchers, and anyone working with evolving data sets [5], [6].

In the following sections, we will first discuss related work in the field of data drift detection and data versioning, then outline our methodology for this study, followed by the results of our experiments, and finally conclude with a discussion of the implications of our findings and suggestions for future research directions.

II. Related work

Storage-orientated approaches to data versioning in large repositories have mainly focused on reducing data duplication and improving storage techniques in the repository itself[7], [8], rather than intelligently detecting data differences, that is, data drift i) *virtual* where the distribution of the dataset changes over time, but the underlying concept remains—-the mapping from features to labels–remains; or ii) *conceptual* where the underlying abstraction changes [9]. In highly dynamic data-intensive environments, datasets can evolve forming a continuous between virtual and conceptual data drift [10].

Although there are certain methods for identifying data drift, there are some limitations despite its significant implications for ML systems [11]. Traditional statistical hypothesis testing methods are among the most commonly used, with techniques such as the Kolmogorov-Smirnov test and Kullback-Leibler divergence offering ways to quantify changes in distributions [12], [13]. However, these approaches often require large samples and can be computationally intensive, making them less suitable for real-time detection. In contrast, ML-based approaches focus on using models to capture the underlying structure of the data and then monitoring for deviations from this structure. Techniques such as Principal Component Analysis (PCA) [14] and Autoencoders (AEs) [15] can capture complex patterns and adapt to high-dimensional data. More recent attempts include software libraries with methods to measure the distance between data distributions [16].

On the other hand, existing work on data versioning tends to focus on version control systems that can track changes but often lack the granularity needed to manage evolving datasets. Despite this, there is a growing need for a standardised approach to data versioning that incorporates data drift metrics and supports a more comprehensive understanding of how data evolve over time [5], [6], [2].

A. Contribution

> In the context of AI and scientific communication, the implementation of FAIR principles, particularly in data versioning, is critical to ensuring that datasets are not only accessible and reusable, but also robust enough to support the evolving requirements of AI technologies. This study, by exploring the quantification and management of data drift, directly supports the development of more reliable AI applications by ensuring data consistency over time, which is crucial for the credibility and precision of scientific communications.

Our initial research has introduced three metrics to measure data drift using a combination of PCA and AEs [17], then extended into a proof-of-concept study to demonstrate the potential for a standardised FAIR-compliant data versioning framework incorporating these metrics for time series [3], an area where data drift has received significant attention due to the increase in data streaming techniques [18]. However, significant questions remain about the robustness and scalability of these methods, especially in real-world scenarios where the size, noise, and other factors of the datasets can vary.

The aim of this paper is therefore to address these questions by exploring the limitations and performance of these data drift metrics under different conditions using a significant number of datasets from a large open data repository, the European data portal. We seek to provide a more robust and adaptable approach to detecting data drift and a clearer path toward standardised dataset versioning.

III. Methodology

This study investigates data drift metrics to understand their behaviour in various experimental setups. We built on previous work that proposed data drift metrics based on PCA and AE. Our focus here is on the two data drift metrics derived from the reconstruction error in these models: $d_{E,PCA}$ and $d_{E,AE}$.

We examine the robustness and reliability of these metrics through experiments that simulate different types of data evolution events: creation, update, and deletion. This allows us to explore the impact of dataset size, noise, and other factors on data drift metrics, as well as to assess computational efficiency and repeatability.

A. Accessing datasets from the European Data Portal

The datasets used in our experiments are sourced from the European Data Portal and are predominantly tabular in nature, stored in CSV format. Each dataset consists primarily of numeric data, which is crucial for the application of PCA and Autoencoders, as these models require quantitative inputs for the computation of metrics like mean squared error and reconstruction error. In addition, the data sets are multidimensional, containing multiple features that record various aspects of the data subject matter.

To access and obtain datasets from the European Data Portal [19], we used a customised Python script to query the API of the portal, filter the results for relevant datasets, and download the data for further analysis. The following subsections describe the methodology used to acquire and process these datasets.

1) *Data acquisition*. We used an API endpoint to acquire datasets from the European Data Portal to allow for structured queries based on various parameters such as search terms, format, and scoring. The key steps in this process included i) Query Generation. A query string was generated with specific parameters to identify datasets available in CSV format. Our research has focused on datasets with good scoring (above a certain threshold) and also matching a search term (in this case, "time_series"); ii) Data Retrieval: We used the requests library to send HTTP GET requests to the API endpoint, passing the generated query string described above as a parameter. This allows us to retrieve information, as metadata, about the datasets that match the search criteria; and, iii) Metadata Storage: The metadata from

the API response are stored in JSON format for future reference. In addition, the acquired metadata include information about the datasets such as their scoring, identifiers, country of provenance, and data distributions.

2) *Dataset filtering and download*. After retrieving the metadata for the datasets, we applied additional filters to ensure that the downloaded datasets met specific criteria i) Scoring Threshold: According to the API documentation, there are four scoring categories. The scoring value of 221 or above represents the minimum threshold imposed by the first two superior scoring categories, ensuring high-quality datasets; ii) Dataset Format: The script checked the format of each dataset, keeping only the CSV files to maintain consistency in data processing; and, iii) Download Datasets: For each data set that met the criteria, we obtained the access URL from the metadata and downloaded the CSV files using the "get()" method from the requests library. Following our interest in generating a unique identifier for the datasets, we title them using the Secure Hash Algorithm 3 (SHA-3) algorithm that provides high collision resistance. Moreover, the dataset download process was run in a separate thread to allow parallel processing and improved efficiency. This allowed the script to process multiple datasets simultaneously, reducing the overall download time.
3) *Validation and Processing*. After downloading the datasets, we applied validation and processing steps to ensure data quality and compliance with our research objectives. i) File Validation: Specific methods were used to validate that the downloaded datasets were indeed CSV files. If a file was corrupted or incompatible, it was discarded; ii) Time Series Identification: In this stage, we checked whether the datasets contained valid date formats to identify potential time series data. This involved matching date patterns to ensure that the datasets were suitable for our analysis; iii) Data Processing and Reorganisation: The validated datasets were processed to obtain a sample of four data points, randomly selected to check for date compliance. Datasets that contained valid time series data were moved to a separate directory for further analysis.

Following this approach, we systematically retrieved, validated, and processed a total of 223 datasets from the European Data Portal. The final result was a collection of CSV files containing time series data, ready for further analysis and experiments, and ensuring a minimum level of quality to save memory space and time that should have been used otherwise in the latter stages to check the usability of the downloaded datasets. In terms of the organisation of the time series within these CSV files, each row corresponds to a time-stamped record, with columns representing different variables measured at each time point. This structure allows us to apply systematically data drift detection techniques by analysing changes in time steps, making the data particularly suitable for our study on the robustness of data drift metrics.

However, harmonising the dataset models was still a further step that resulted in a benchmark of 71 usable datasets for our analysis displayed in Section IV. Yet, despite using a reduced amount of datasets compared to the number and wide range of time series available in a repository such as the European Data Portal, our results illustrate a potential harmonisation solution to work with different time series datasets. The methodology presented in this paper arguably provides a robust framework for accessing and working with large-scale datasets from online repositories, ensuring consistency and quality throughout the data drift calculation, paving the way toward automatic standardised dataset versioning with FAIR compliance.

B. Data Drift Metrics

Our approach to quantifying data drift uses unsupervised ML models, focussing on PCA and AE. These models offer flexibility in analysing high-dimensional data without predefined prediction targets. In the context of our experiments, the term "*cells*" refers to the individual data points within the data set matrix, where each cell represents a unique intersection of a row and a column, that is, a specific numeric value at a given time step for a particular variable. To assess the sensitivity of our data drift metrics, we employed a permutation strategy designed to disrupt the natural correlation patterns within the dataset. Specifically, a designated percentage of cells, denoted by $p\%$, within each variable is randomly permuted.

This permutation is performed independently for each column, ensuring that the intrinsic structure of each variable is altered to a predefined extent. This method of permutation results in a new dataset variant where the original order of data points in each column is disrupted, simulating a scenario of potential data corruption or unanticipated changes in data collection procedures. The rationale behind using this permutation process is to evaluate how well the PCA and AE models, which have been trained on the original unpermuted dataset, can adapt to or recognise changes in the underlying structure of the data set. The increase in Mean Squared Error (MSE) during model reconstruction of the permuted dataset provides a quantitative measure of the model's sensitivity to the induced data alterations, reflecting the robustness of the data drift metrics under conditions of simulated data drift. The following sections explained how we combined this strategy with the two aforementioned unsupervised ML models.

1) E-drift Metric based on PCA: The PCA-based E drift metric ($d_{E,PCA}$) measures drift by assessing the Mean Squared Error (MSE) between the reconstructed data and the original data, using a PCA model. This metric is obtained through a permutation strategy, in which different percentages of cells are permuted from the primary source dataset (PS), ranging from 1% to 100%, with increments of 10%. Each level of permutation is repeated several times (typically 10 times) to ensure robustness. The resulting MSE values are used to fit splines that represent the relationship between the MSE and the percentage of permutation, with a higher value indicating greater data drift.

2) E-drift Metrics based on Autoencoders: Similarly, the AE-based E-drift metric ($d_{E,AE}$) uses an AE to compute the reconstruction error. The same permutation strategy is applied, with MSE values derived from the reconstruction process. The splines are fitted to determine the estimated percentage of permutation that corresponds to the observed MSE. This allows us to obtain a metric that ranges from 0 to 100, indicating the level of drift between the PS dataset and the Revision (R) dataset.

> It is important to mention that alongside the data drift metrics, metrics $t_{E,PCA}$ and $t_{E,AE}$, referring to the computational times required to compute the PCA and AE models, respectively, have also been calculated, reflecting both the complexity of the model calculations and the overhead associated with handling larger or more complex datasets.

C. Experimental Setup

To evaluate the robustness of these data drift metrics, we have performed experiments that simulate three types of data evolution event: creation, update, and deletion. Each experiment was repeated multiple times to ensure consistency and assess the impact of different variables on the metrics' behaviour.

1) Creation Event: The creation event involved adding new information to the PS dataset, simulating a dynamic dataset scenario. We generated new observations using overlapping windows of 5%, 10%, 25%, 50%, 75%, and 100% of the remaining records in the Revision datasets, which had not been seen to fit the PCA and AE models. With each iteration, new batches of new records were added, while removing the first record from the previous iteration. This allowed us to explore how varying time resolutions and memory sizes impact the data drift metrics. To create new observations for our experiments, we simulated a scenario in which data continuously accrue over time, mimicking real-world incremental data collection. In practice, this approach involved presenting subsets of data that had not been used to train the ML models, in overlapping sequences. For example, with a window size of $n = 2$, the procedure would start by introducing the first and second records of the subset to the model, followed by the second and third records, etc. This overlapping ensures that each new data window introduced to the model differs slightly from the previous one by precisely one observation. This is particularly effective in highlighting how well data drift metrics respond to subtle shifts in data patterns over successive time points. This technique not only tests the sensitivity of the models to new information, but also simulates a practical environment where data drift might naturally occur due to ongoing data accumulation.

2) Update Event: The update event involved changing existing records in the dataset by modifying the scale of the variables. We applied cubic root transformations to different percentages of columns (5%, 10%, 30%, 50%, 70%, 80%, and 100%) to simulate changes in the underlying data structure. This approach aimed to assess the sensitivity of the data drift metrics to such transformations and to assess whether substantial updates would trigger significant data drift values.

3) Deletion Event: The deletion event simulated data loss by decimating signals and removing different percentages of records from the PS dataset (5%, 10%, 20%, 30%, 40% and 50%). This allowed us to assess the resilience of unsupervised ML models when faced with incomplete data and to determine how information loss affects data drift metrics.

D. Data Preparation and Preprocessing

The first step before executing the Creation, Update, and Deletion experiments, was to preprocess and fit the PCA and AE models with the Primary Source datasets. In practical terms, each data set $\mathbf{X}$ is considered a matrix with dimensions $N \times K$, where N represents the number of time steps (records) and K is the number of variables measured at each time point. Therefore, each record $\mathbf{x}_t$ within the dataset constitutes a K-dimensional observation vector, capturing the values of the K variables at a specific time point t. For the purposes of our experiments, the datasets were partitioned into training and testing sets to evaluate the robustness and predictive capability of our models in Creation experiments. Specifically, each dataset was divided so that the first half of the records was used to fit the PCA and AE models (training set, or Primary Source), and the second half was reserved for testing the models' performance in detecting drift (testing set, or Revision). This setup ensures that each record, or row in the dataset matrix, is treated as a separate, complete observation for model training and testing, reflecting the typical structure of time series data where each time step is crucial for understanding temporal dynamics.

It is important to mention that prior to fitting the models the experiments, each feature in the dataset was scaled individually to have zero mean and unit variance, a process often referred to as standardisation or autoscaling. This ensures that all features contribute equally to the analysis, preventing features of larger scale from dominating the results.

For PCA, the models were fitted to explain more than 90% of the variance. However, AEs require a more sophisticated training routine. The ones used in our study are fully connected deep neural networks with two main components: an encoder and a decoder, both composed of two hidden layers and a variable number of nodes. The activation functions used were "relu" and "tanh" for the hidden layers, with linear activation for the final output layer. This architecture aims to capture complex non-linear patterns within the data for effective data drift detection by compressing and reconstructing the input data. This reconstruction error was used later as a measure of data drift. Noise was added to the data to improve the AE's generalisation ability. This noise followed a normal distribution with a mean of zero and a standard deviation scaled by a factor ($c_\epsilon = 0.005$), applied after autoscaling the data. The AE was trained using a customised training routine, including hyperparameter tuning through Keras Tuner,

TABLE I: Features summarising the performance of the data drift metrics.

Feature	Meaning
N	Number of records (time steps, rows) of the dataset
K	Number of records (time steps, rows) of the dataset
R^2_{PCA}	Goodness-of-fit (i.e.: explained data variance) of the PCA model fit with the Primary Source dataset
R^2_{AE}	Goodness-of-fit (i.e.: explained data variance) of the AE model fit with the Primary Source dataset
Rd_{PCA}	Range of average $d_{E,PCA}$ across all levels of the experiment
Rd_{AE}	Range of average $d_{E,AE}$ across all levels of the experiment
Σd_{PCA}	Addition of average $d_{E,PCA}$ values for each experiment level, normalised by the number of levels (ensuring $\in [0, 100]$)
Σd_{AE}	Addition of average $d_{E,AE}$ values for each experiment level, normalised by the number of levels (ensuring $\in [0, 100]$)
Δd_{AE-PCA}	Maximum value of the difference between the average values $d_{E,PCA}$ and $d_{E,AE}$ of each level of experiment.
Δt_{AE-PCA}	Maximum value of the difference between the average values $t_{E,PCA}$ and $t_{E,AE}$ of each level of experiment.

optimising for the lowest validation mean squared error. Early stopping was used to prevent overfitting, and training stopped if the validation loss did not improve for 10 consecutive epochs. This approach not only enhances the generalisability of the model, but also optimises computational efficiency.

After obtaining the values $d_{E,PCA}$, $d_{E,AE}$, $t_{E,PCA}$ and $t_{E,AE}$ for each iteration of each level of the experiments, the results were aggregated. This was done to achieve a comparable structure of information that encapsulates the results obtained for each dataset. To do so, we had to do some feature engineering and selection, picking the optimal parameters summarising the data drift and computation time values obtained at each level of the experiments performed in each versioning scenario. In this particular case, optimality was defined by the insight brought by the parameters but also by the level of variability. The existence of some outliers forced us to perform the latter unsupervised analysis using robust techniques that prevented the generation of artefact clusters due to the existence of anomalous points. With this goal in mind, a summary dataset was created, containing the following information for each dataset: the number of rows of the Primary Source (N), the number of features of the Primary Source (K), the goodness of fit of the PCA model for PS (R^2_{PCA}), AE goodness-of-fit for the PS (R^2_{AE}), and several metrics expressing the variation of and between $d_{E,PCA}$, $d_{E,AE}$, $t_{E,PCA}$, and $t_{E,AE}$, across all levels of experiments (see Table I). Once the results were aggregated, these summary datasets were preprocessed to ensure consistency and comparability of the experiments.

E. Unsupervised analysis

This work aims to critically assess the performance of $d_{E,PCA}$ and $d_{E,AE}$ metrics and determine their applicability in real-world scenarios where datasets evolve and differ in size, structure, and scale. This means providing a quantitative foundation for evaluating the stability and reliability of the proposed metrics and offering a more systematic approach to understanding data drift in diverse datasets. Given the expanded range of data sets used in this study compared to our previous work, robust PCA [20], [21] was used to analyse the summary data sets derived from the creation, update and deletion experiments. This technique allowed us to explore relationships between datasets showing similar behaviours and to interpret the underlying variables correlated to these groupings of datasets, establishing potential relationships between dataset characteristics and data drift metrics, as well as computational times.

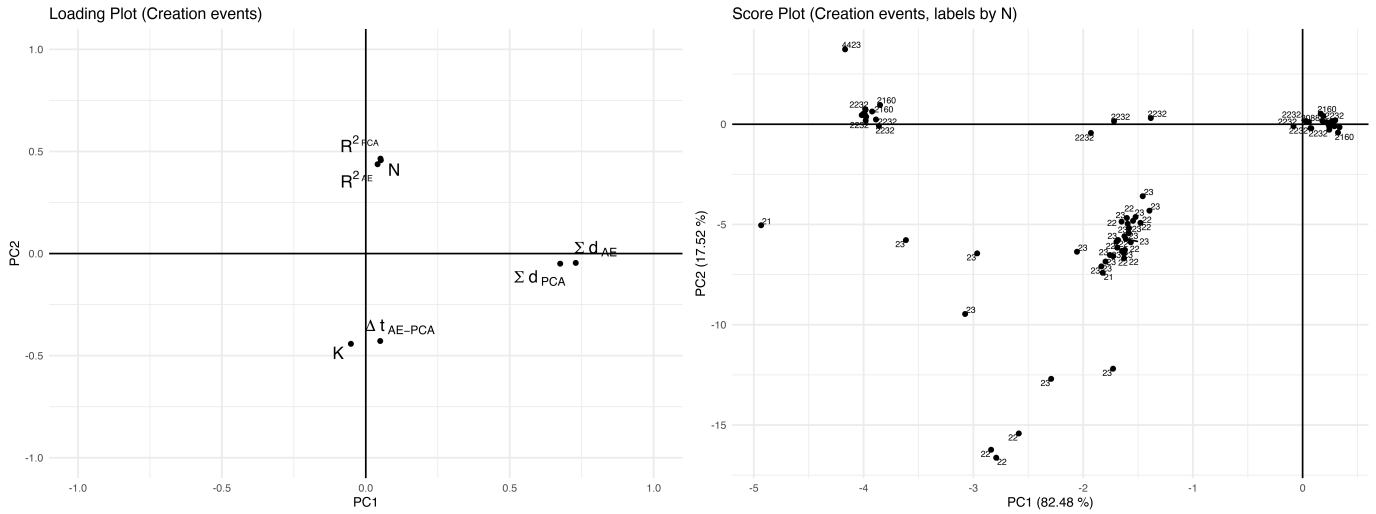
IV. Results

The following figures depict the results of the PCA run on the matrices that aggregate the results for the Creation, Update, and Deletion events. Two different plots will be used to inform the results obtained, presenting two distinct visualisations. On the one hand, the loading plot serves as a map to interpret the variables' correlations, and it will be used to showcase the contributions of each variable to the first two principal components. Each point represents a variable, with its position indicating the relative contribution to the PCA axes. Score plots represent correlation maps between observations, where each point corresponds to an observation (i.e. a dataset in this case), positioned according to its coordinates on the first two principal components.

A. Creation events

Loading plots shown in Figure 1a, show that the most significant variance in PC1 is associated with $\Sigma d_{E,AE}$ and $\Sigma d_{E,PCA}$ (i.e., the normalised sum of the average data drift values from the creation experiments). This means that in these experiments, both metrics show a major alignment. In addition, their high values on the 1st PC loadings and close to zero values for the 2nd PC indicate orthogonality between the data drift metrics and the rest of the variables, more related to the datasets' characteristics. This suggests that data drift values, in this case, agree for all types of dataset, regardless of their features.

The loadings of the second PC indicate that the variance in PC2 is related to the goodness of fit of the models, the sample size of the dataset (N), and the number of variables. The alignment of the positive PC2 values of both R^2 for PCA and AE and N indicates that datasets with more time steps will present better quality than the primary source dataset, bringing more information to fit both models. In addition, the negative loading of K (number of characteristics) indicates that achieving better R^2 values becomes equally challenging for both PCA and AE models, since their goodness of fit is negatively correlated with the dimensionality of the dataset



(a) Loading plot with the 1st and 2nd PCs on the x and y-axis, respectively. (b) Score plot with the 1st PC on the x-axis and of the 2nd PC on the y-axis.
Fig. 1: First two components of the robust PCA model obtained with the Creation events experiments. The plot on the left represents the loading plot, which illustrates the contributions of variables to these two main components. The right plot is the score plot, with each point representing an observation, labelled by its corresponding value of variable N, indicating the number of time steps (rows) of each dataset.

K. Finally, it is interesting to mention that the number of ov variables K seems to be positively correlated with longer computation times by the AEs compared to the computation time required by the PCA models.

The score plot in Figure 1b shows that, in fact, there are three more differentiated groups of datasets whose dispersion along the 2nd PC correlates with lower N values. However, as mentioned before, the cluster located on the left side of the plot still maintains high values of N of a similar order of magnitude to the cluster located close to the centre of the score plot. The labels are shown in Figure 2, which shows the normalised values d_{PCA} in all creation event experiments, indicating that this division of clusters along the 1st PC corresponds to different levels of variability captured in the Revision datasets of the experiments. Yet, as mentioned before, the variables' loadings indicate that such differences in the data drift values do not respond to the limitations of the proposed data drift metrics when working with datasets of certain characteristics. Thus, it seems reasonable to assume that revisions from datasets located in the left cluster presented lower drifts from the Primary Source than the ones located at the centre of the score plot.

It is worth mentioning that these findings also suggest that ML models, PCA and AE, seem to face the same limitation when fitting datasets of similar characteristics. According to these results, low N and high K data sets could threaten the reliability of data drift metrics. Further insight on the sensitivity of these metrics should be done, assessing their limitations and ideally proposing other candidates to deal with datasets with such characteristics.

B. Update events

The loading plots in Figure 3a reveal that PC1 is mainly associated with the maximum differences in data drift metrics

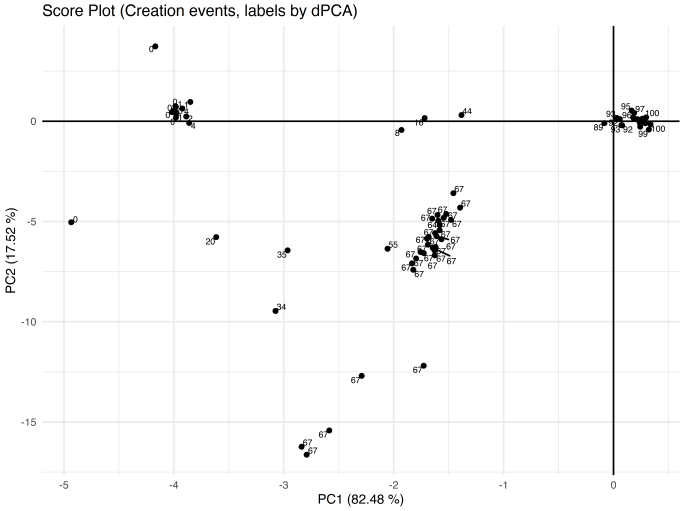
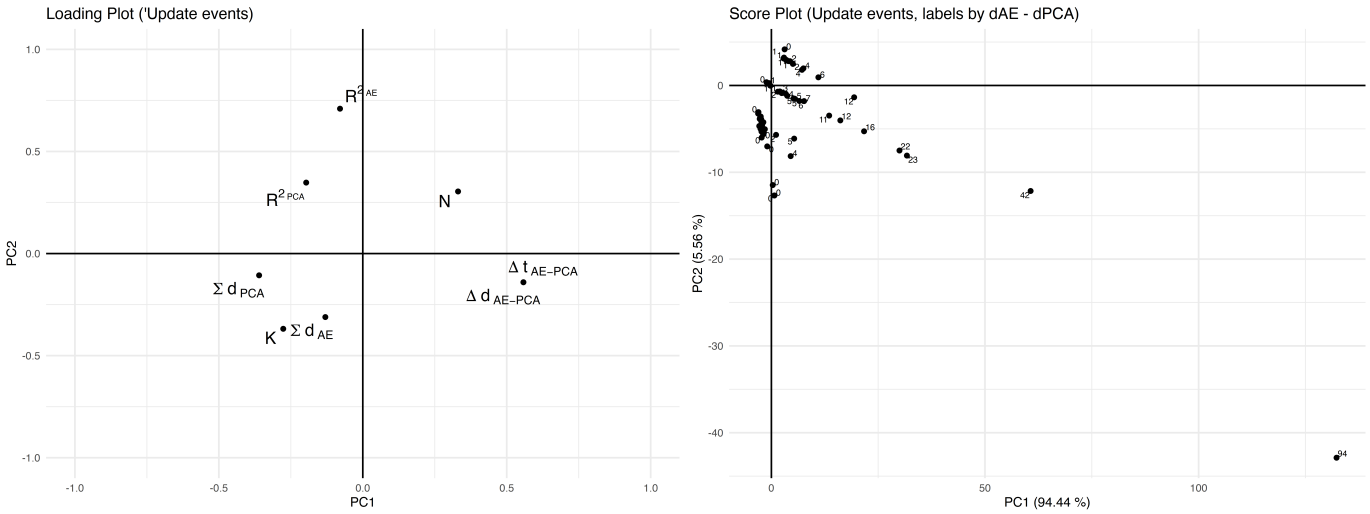


Fig. 2: Score plot for the creation events results.

and computation times between AEs and PCA. The high loadings for the maximum differences between the average data drift values and their average computation times (Δd_{AE-PCA} and Δt_{AE-PCA}, respectively) suggest that PC1 is driven by the point of greatest disagreement between these two metrics, indicating that data sets with large discrepancies in data drift and computation times dominate this principal component. Furthermore, the positive loading for the number of records N and the negative loading for the number of features K, also suggest that data sets with more time steps but fewer features may exhibit larger discrepancies between AE and PCA. However, this last remark could be explained by the simple fact that datasets with lower dimensionality (i.e., lower K), will be more sensitive to shifts in the scale of their



(a) Loading plot with the 1st and 2nd PCs on the x and y-axis, respectively. (b) Score plot with the 1st PC on the x-axis and of the 2nd PC on the y-axis.
Fig. 3: First two components of the robust PCA model obtained with the Update events experiments. The left plot represents the loading plot, illustrating the contributions of variables to these two principal components. The right plot is the score plot, with each point representing an observation, labeled by its corresponding value of variable Δd_{AE-PCA}, indicating the maximum difference between the average data drift values obtained by each approach at each experiment level.

variables, as they will have a higher impact on the overall correlation pattern described by the new shifted version of the dataset.

Furthermore, observing PC2, it can be seen that the AE's goodness of fit (R^2_{AE}) is again negatively correlated with the dimensionality of the dataset K. However, in this case, even if PCA's goodness of fit (R^2_{PCA}) and the number of records (N) are still positively correlated to R^2_{AE} according to PC2, AEs seem to be more affected by the dimensionality of the dataset, as both R^2_{PCA} and N present loadings closer to zero in this case, in comparison to the relationship exhibited in the loading plot from Creation experiments (Figure 3a). The combination with other variables in this analysis appears to widen the sensitivity of AE and PCA to the dimensionality of the dataset, making the latter slightly more robust.

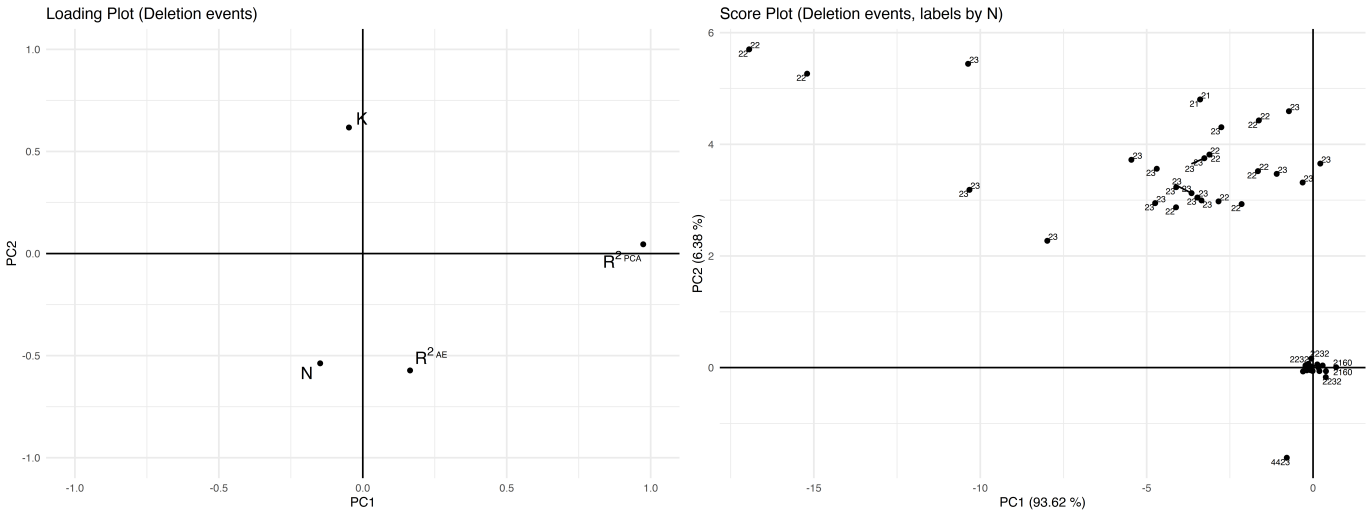
In addition, loadings on the second PC also seem to indicate that R^2_{AE} is negatively correlated with the normalised sum of $d_{E,AE}$ values, which means that the datasets whose Primary Source was modelled more accurately presented lower $d_{E,AE}$ values. On the other hand, the first PC expresses an antagonistic dynamic for PCA models. The negative loadings of R^2_{PCA} and $d_{E,PCA}$ indicate that the data sets whose primary source was modelled more accurately presented higher values of $d_{E,PCA}$. The reason behind this might be the difference in each model's linear and nonlinear nature. Since AEs can model non-linear relationships, performing non-linear scale shifts on the variables might not distort completely the captured correlations. In contrast, since PCA relies on linearity assumptions when linear relationships are broken by inducing non-linear transformations, $d_{E,PCA}$ might be more sensitive towards them. This would mean that AEs might be more flexible to certain data shifts, which could be beneficial if such changes are expected on the natural dynamics of datasets, but could be counterproductive otherwise, making the $d_{E,AE}$ metric less sensitive to the changes in these new versions.

Finally, datasets with fewer features but more time steps seem to yield greater discrepancies between both data drift metrics, both in terms of their values and in computation times. This, alongside the aforementioned results, suggests that update events, such as scale shifts, highlight the differences between both data drift metrics, pointing out a direction to conduct further assessment and an in-depth analysis of the particular cases that could be problematic for each approach.

The score plot in Figure 3b indicates that there are distinguishable groups, but the presence of a dataset with a difference between the average values $d_{E,AE}$ and $d_{E,PCA}$ of $\Delta d_{AE-PCA} = 94$, is probably the most noticeable aspect. This dataset represents an outlying behaviour that combines a high dimensionality ($K = 48$) and a bad AE fitting for the Primary Source ($R^2_{AE} = 0.3646$). This outlying case illustrates a scenario that could be challenging for the use of $d_{E,AE}$'. Consistency in model goodness-of-fit is essential for ensuring the reliability of data drift metrics, and further analysis should focus on assessing the underlying factors causing discrepancies and exploring ways to maintain reliable data drift metrics even when handling datasets with potential scale shifts.

C. Deletion events

The first thing to mention about the deletion experiments is that the results presented such low variability that the robust PCA model considered their elimination to carry out the analysis, given the low variability they presented. Even if striking, this result can be better understood if the results



(a) Loading plot with the 1st and 2nd PCs on the x and y-axis, respectively. (b) Score plot with the 1st PC on the x-axis and of the 2nd PC on the y-axis.
Fig. 4: First two components of the robust PCA model obtained with the Deletion events experiments. The left plot represents the loading plot, illustrating the contributions of variables to these two principal components. The right plot is the score plot, with each point representing an observation, labeled by its corresponding value of variable N, indicating each dataset's number of features (columns).

of our previous work are checked. In our previous proof-of-concept results, both $d_{E,PCA}$ and $d_{E,AE}$ metrics showed high robustness against information loss, with values of null data drift almost persistently for all percentages of deleted information from the Primary Source.

For this reason, the loading plot in Figure 4a shows only the correlation between the dimensions of the dataset and the goodness-of-fit of both ML models fitted to the primary sources. However, this case is also interesting to analyse. The high loading for R^2_{PCA} on the first PC and its orthogonality to the rest of the variables (with values close to zero for the first PC) implies that the goodness of fit of the PCA models is the main source of variability in the results of the deletion experiments and that this variability within R^2_{PCA} does not appear to be strongly correlated with the dimensionality of the data sets or follow the same pattern as the goodness of fit obtained by the AEs (informed by R^2_{AE}). In contrast, this does not seem to be the case for R^2_{AE} values, which appear to be more sensitive to these characteristics of the dataset, increasing their reliability with the number of records (positive correlation with N, aligned in the second PC) and decreasing with higher dimensions (negative correlation with K according to the 2nd PC).

However, the loadings of the 1st component clearly attribute higher variability among R^2_{PCA} values. This has a direct implication on the reliability of PCA models, which would be expected to be more variable than that of AE models. Even if this variability cannot be attributed to the dimensions of the dataset, it might be interesting to further assess which dataset's features might correlate with variations on PCA's performance. In practical terms, this means that, in the case of using the $d_{E,PCA}$ metric, assessing the Primary Source's goodness-of-fit would be critical, regardless of the dataset's features, and, on the other hand, using the $d_{E,AE}$ metric with datasets of a small number of records and high dimensionality could weaken the reliability of the $d_{E,AE}$ metric.

D. General discussion

The results compared the performance of two candidates for data drift based on two different unsupervised ML models. The difference in handling non-linear shifts between PCA and AE is predominantly due to their inherent mathematical constructions. As a linear technique, PCA is primarily effective in capturing the major linear variances within the data, but it fails to model nonlinear relationships effectively.

In contrast, AEs are designed to capture non-linear dependencies thanks to their layered neural network-based structure. Their performance in high-dimensional settings is significantly affected due to their complexity and the higher risk of overfitting. As the dimensionality of the dataset increases, the AE must learn more parameters, which can lead to model overfitting unless sufficient regularisation strategies are employed. This sensitivity to high dimensionality can be attributed to the model's capacity to capture intricate patterns, which, while beneficial for capturing complex nonlinear relationships, also makes the model prone to fitting noise present in the data. To mitigate these issues, techniques such as dropout, reduction in dimension before training, and increasing the amount of training data can be considered. Each of these strategies helps reduce the effective complexity of the model or the noise in the training data, leading to stronger performance in high-dimensional scenarios.

Practically, this implies that PCA might be more suitable for datasets where the underlying data relationships are linear or approximately linear due to its simplicity and less computational cost. However, AEs are preferable in scenarios

where data exhibit complex non-linear patterns, albeit at the cost of increased computational resources and a higher propensity for overfitting. In choosing between these methods, practitioners should consider the nature of the dataset and the specific requirements of their application, such as the need to capture non-linear relationships versus the computational efficiency and simplicity of the model. It is also advisable to employ techniques such as cross-validation to evaluate the performance of each model on the dataset before finalizing the approach.

The implications of our findings extend beyond the mere characterisation of data drift. Our experiments highlight the need for robust metrics that can adapt to various types of data change, ensuring the consistency and reliability of the data over time. By systematically evaluating how data drift metrics respond under different creation, update, and deletion conditions, we provide a foundational understanding that can guide the development of standards for dataset versioning practices, which are integral to implementing FAIR principles across diverse data ecosystems. We envision that adopting such standardised practices will lead to better governance of data assets across various sectors. To this end, the framework we depict should include guidelines to i) Assess the stability of data drift metrics across different dataset characteristics and update cycles; ii) Implement automated tools that use these metrics to flag significant deviations or drifts in data, prompting re-evaluation of data models; iii) Develop a comprehensive metadata management system that documents each versioning event along with the associated drift metrics, facilitating easier traceability and accountability in data management.

These recommendations are designed to foster a more standardised approach to handling evolving datasets, thus contributing to the integrity and utility of data in AI applications and beyond.

V. Conclusion

In this work, we have assessed the reliability and variability of the data drift metrics for a FAIR-compliant standard for data versioning. These metrics are derived from PCA and AE, and we have evaluated their performance by simulating creation, update, and deletion events. Our approach used robust PCA to analyse the aggregated results to identify significant patterns, trends, and outliers and inform best practices for data drift detection.

Firstly, analysis of creation events revealed that both data drift metrics ($d_{E,AE}$, $d_{E,PCA}$) showed a major alignment, indicating consistency between different datasets. The PCA loading plot showed a positive correlation between the normalised sum of the average data drift values, suggesting that these metrics coherently represent the underlying drift in these experiments. In addition, our results also suggest that datasets with more time steps generally exhibit better model fitting, providing a more stable baseline for detecting data drift. However, both PCA and AE models struggle with higher dimensionality, indicating a challenge for datasets with more features being measured.

Secondly, update events highlighted potential discrepancies between the two data drift metrics and their computation times. The maximum differences between the average data drift values and their computation times are relevant sources of variability, suggesting substantial differences in these metrics between data sets. This may indicate that update events can reveal deeper complexities and sensitivities within these data drift metrics. In addition, AE's goodness-of-fit presented a strong negative correlation with the dataset's dimensionality, suggesting that higher dimensionality may pose challenges to the AE's stability and accuracy. In contrast, PCA models seem to be more robust in these cases, demonstrating a more consistent behaviour when faced with nonlinear shifts or variations in dataset scale.

Finally, deletion events show limited variability in the results, leading to a robust PCA analysis that ignores all comparative metrics between data drift values or execution times with AEs and PCA models. This result aligns with earlier findings that both data drift metrics demonstrated robustness against information loss. However, the resulting PCA model still indicated higher variability within the PCA model goodness-of-fit, suggesting that accurate PCA fitting might not always be ensured and that the reasons behind this variability did not seem strongly correlated with the dataset dimensions. However, the analysis also suggests that smaller datasets present a greater challenge to the reliability of the PCA model, pointing to potential problems with limited data. However, correlations of N and K with AE's goodness-of-fit suggest a stronger impact of dimensionality on $d_{E,PCA}$ reliability, reinforcing the need for caution when dealing with high-dimensional, low-sample datasets.

> In summary, these results highlight the performance of the data drift metrics derived from PCA and AE. Although they can be effective in capturing underlying data shifts, they are sensitive to certain dataset characteristics and model fitting issues. Datasets with lower dimensionality and higher time steps tend to produce more consistent data drift metrics, while those with higher dimensionality pose challenges, especially for AEs. The goodness of fit for both the PCA and AE models plays a crucial role in determining the stability of the data drift metrics, indicating that careful model fitting and tuning are essential for reliable results.

Our findings underscore the importance of integrating FAIR principles in dataset versioning to enhance the reliability and applicability of AI models in scientific communications. By ensuring that data drift is accurately measured and managed, we contribute to the broader goal of making AI technologies more adaptable and effective in handling real-world data variations, presenting a clear pathway towards more standardised and robust scientific tools. Given these findings, future work will focus on refining data drift metrics to address identified

limitations, such as sensitivity to the dimensionality of the dataset and model-fitting challenges. Based on the results presented, several potential avenues for future research such as i) *Hybrid Metrics:* Given the strengths and weaknesses of PCA and AE observed in handling different data characteristics, a hybrid approach could be explored combining PCA efficiency with AE's non-linear modelling capacity. Using ensemble methods that aggregate the results of multiple drift detection models could also improve the stability and accuracy of drift measurements. However, while this approach might help mitigate the specific shortcomings of individual models, it would also raise more questions about how to interpret the contradictory conclusions obtained by each metric; ii) *Application of Deep Learning Techniques:* Exploring deep learning architectures, such as Attentional or Recurrent Neural Networks, which are well suited for pattern recognition in time-series data, could provide new insights into more effective ways of detecting and quantifying data drift, especially in dynamically changing environments; and, iii) *Real-time Drift Detection Algorithms:* Developing real-time data drift detection algorithms should be highly beneficial, particularly for applications requiring immediate response, such as dynamic pricing models or real-time systems monitoring.

These suggestions are designed not only to address the limitations found in our current metrics but also to push the boundaries of research in data drift detection, ensuring that data management systems remain robust against the evolving nature of data environments. This is particularly interesting in the context of our current SMARDY research project [22], which is developing a marketplace for research datasets where the detection of data drift has proven to be crucial.

Acknowledgments

This work has been developed under the auspices of the European research project "SMARDY: Marketplace for Technology Transfer of Research Data, Software, and Results" (https://smardy-project.eu/) funded from 2021 to 2024 by the European Eureka Network programme through i) Ireland's International Research Fund of Enterprise Ireland (Ref#: IR20210058); and ii) Romania's Ministry of Research, Innovation, and Digitalisation CNCS/CCDI-UEFISCDI (Ref#: PN-III-P3-3.5-EUK-2019-0241) PNCDI III.

References

[1] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[2] J. Klump, L. Wyborn, M. Wu, J. Martin, R. R. Downs, and A. Asmi, "Versioning data is about more than revisions: A conceptual framework and proposed principles," *Data Science Journal*, vol. 20, no. 12, pp. 1—-13, Mar 2021.

[3] A. González-Cebrián, M. Bradford, A. E. Chis, and H. González-Vélez, "Standardised versioning of datasets: a FAIR–compliant proposal," *Scientific Data*, vol. 11, no. 358, pp. 1–15, 2024.

[4] M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 160018, pp. 1–9, 2016.

[5] A. Treloar, "The Research Data Alliance: globally co-ordinated action against barriers to data publishing and sharing," *Learned Publishing*, vol. 27, no. 5, pp. S9–S13, 2014.

[6] D. B. Allison, A. W. Brown, B. J. George, and K. A. Kaiser, "Reproducibility: A tragedy of errors," *Nature*, vol. 530, pp. 27–29, 2 2016.

[7] S. Bhattacherjee, A. Chavan, S. Huang, A. Deshpande, and A. Parameswaran, "Principles of dataset versioning: Exploring the recreation/storage tradeoff," *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1346–1357, 2015.

[8] A. P. Bhardwaj, S. Bhattacherjee, A. Chavan, A. Deshpande, A. J. Elmore, S. Madden, and A. G. Parameswaran, "Datahub: Collaborative data science & dataset version management at scale," in *Conference on Innovative Data Systems Research (CIRDB) 2015*. Asilomar: CIRDB, Jan. 2015.

[9] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Advances in Artificial Intelligence, SBIA 2004*, ser. LNCS, vol. 3171. Maranhao: Springer, Sep. 2004, pp. 286–295.

[10] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, mar 2014.

[11] S. Rabanser, S. Günnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Advances in Neural Information Processing Systems, NeurIPS 2019*, vol. 32. Vancouver: Curran, Dec. 2019.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[13] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[14] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37–52, 1987.

[15] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[16] J. Céspedes Sisniega and A. López García, "Frouros: An open-source Python library for drift detection in machine learning systems," *SoftwareX*, vol. 26, p. 101733, 2024.

[17] A. González-Cebrián, L. A. McGuinness, M. Bradford, A. E. Chis, and H. González-Vélez, "Automatic versioning of time series datasets: a FAIR algorithmic approach," in *18th IEEE International Conference on e-Science, e-Science 2022*. Salt Lake City: IEEE, Oct. 2022, pp. 204–213.

[18] A. L. Suárez-Cetrulo, D. Quintana, and A. Cervantes, "A survey on machine learning for recurring concept drifting data streams," *Expert Systems with Applications*, vol. 213, p. 118934, 2023.

[19] "European data portal," accessed: 02 may 2024. [Online]. Available: https://data.europa.eu/en

[20] P. J. R. Mia Hubert and W. Van den Bossche, "MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers," *Technometrics*, vol. 61, no. 4, pp. 459–473, 2019.

[21] T. Verbelen and C. Croux, *cellWise: Analysis of Data with Cell-Wise Outliers*, Comprehensive R Archive Network (CRAN), 2022, version 3.2. [Online]. Available: https://rdrr.io/cran/cellWise/man/MacroPCA.html

[22] I.-D. Filip, C. Ionite, A. González-Cebrián, M. Balanescu, C. Dobre, A. E. Chis, D. Feenan, A.-A. Buga, I.-M. Constantin, G. Suciu, G. V. Iordache, and H. González-Vélez, "SMARDY: Zero-trust FAIR marketplace for research data," in *IEEE BigData*. Osaka: IEEE, Dec. 2022, pp. 1535–1541.