IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"

14-17 February 2022

Anomaly intersection: disentangling data quality and financial stability developments in a scalable way[1]

Gemma Agostoni, ECB, Louis de Charsonville, McKinsey & Company,
Marco D'Errico, ECB, ESRB Secretariat, Cristina Leonte, BIS, and Grzegorz Skrzypczynski, ECB

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Anomaly *intersection*: disentangling data quality and financial (stability) developments in a scalable way

Gemma Agostoni (ECB), Louis de Charsonville (McKinsey & Company), Marco D'Errico (ECB, ESRB Secretariat), Cristina Stefana Leonte (BIS), Grzegorz Skrzypczynski (ECB)

The views expressed in this paper are of the authors and do not necessarily represent the views of the associated institutions.

Abstract

Enhancing the information available to policymakers is one of the pillars of the reforms following the global financial crisis. Several granular data collections have the potential to increase transparency and support effective policy responses. However, poor data quality is impairing this process, making it increasingly arduous to disentangle whether anomalies and risk build-ups are relevant from a policy perspective or the result of poor data quality. We term this problem *anomaly intersection* and propose a general framework to tackle it in a scalable and flexible way. The framework allows to build a set of automatable tools that analyse anomalies at all levels of aggregations, uncovering their nature. We show how this framework can be successfully applied to transaction-level data on derivatives collected under the European Market Infrastructure Regulation, the largest supervisory dataset to date.

Keywords: automatic, granular, derivatives, quality, EMIR, decision trees, financial stability

JEL classification: C18

Contents

1 Introduction

``*Good data and good analysis are the* ***lifeblood of effective*** *surveillance and policy responses''* (FSB, IMF 2009). Very few episodes exemplify this statement more than the global financial crisis and the more recent Covid-19 crisis. Indeed, one of the lessons we keep learning from crises is that prompt access to high quality data is key to develop an effective policy response. Indeed, *''lack of timely, accurate information hinders the ability of policy makers and market participants to develop effective responses''* (FSB, IMF 2009). Enhancing the set of information available to policymakers through the collection of new sources of data has been one of the pillars of the post crisis policy reforms, albeit a lesser studied one. Policy institutions' work is now profoundly connected with the collection and analysis of data. For instance, the task of the European Systemic Risk Board[1] is to ``***monitor and assess systemic risk*** *in normal times for the purpose of mitigating the exposure of the system to the risk of failure of systemic components''* (ESRB Regulation).[2] Such monitoring activities ``*should be based* ***on a broad set of relevant macroeconomic and micro-financial data and indicators***''.

To improve their monitoring and analytical tasks, policymakers are increasingly collecting and analysing an unprecedented wealth of data. ``*Monitoring an interconnected financial system involves the availability of* ***detailed and granular transactions data****.'' (Mario Draghi, 2018)*: [3] indeed, these datasets are collected at a relatively high frequency, and with high level of granularity and details. One of the most well-known granular collections is represented by data reported under the European Market Infrastructure Regulation (EMIR).[4] EMIR mandates entities in the EU to report details of their derivatives contracts to Trade Repositories (TRs), resulting in the collection and processing of about 100 million observations per day. Transaction-level derivatives data represent a particularly relevant instance, given the role the opacity of these instruments played in the amplification and transmission of the Global Financial Crisis. EMIR states that OTC derivatives ``*create a complex web of interdependence which can make it* ***difficult to identify the nature and level of risks involved****. The financial crisis has demonstrated that such characteristics increase uncertainty in times of market stress and, accordingly, pose risks to financial stability''*.

Remarkably, since the inception of the reporting regime, these data have been characterised by substantial, pervasive, and persistent data quality issues. Data quality issues can be traced back to both reporting entities and trade repositories and apply to both large players (including Central Counterparties and large banking groups[5])

[1] The European Systemic Risk Board (ESRB) was established in the aftermath of the global financial crisis, and is responsible for the macroprudential oversight of the EU financial system and the prevention and mitigation of systemic risk.

[2] Regulation (EU) No 1092/2010 of the European Parliament and of the Council of 24 November 2010 on European Union macro-prudential oversight of the financial system and establishing a European Systemic Risk Board (with further amendments)

[3] Welcome remarks at the third annual conference of the ESRB, https://www.ecb.europa.eu/press/key/date/2018/html/ecb.sp180927.en.html

[4] Regulation (EU) No 648/2012 of the European Parliament and of the Council of 4 July 2012 on OTC derivatives, central counterparties and trade repositories (with further amendments)

[5] See also ECB Banking Supervision (2018)

and smaller players.[6] They are very **heterogenous** in nature: from (i) missing reports (that is: a counterparty, including CCPs, not reporting all or part of their contracts), to (ii) limited use of agreed standards, such as the Legal Entity Identifier (LEI) or Unique Trade Identifier (UTI), to (iii) the reporting of values that are implausible, incorrect, not up to date or not reconciled between counterparties entering a contract. The first two root causes, while still rather concerning, can be and are addressed via institutional channels;[7] the latter is left to the data scientist and policy analyst.

We argue that working in the presence of such data quality issues does not represent only a technical and statistical problem:

1. it represents a key hurdle to fully exploiting this wealth of data which, in turn, increases **opacity** for both policymakers and market participants;
2. it hampers the **scalability** of monitoring frameworks for policymakers and
3. it limits the **ability** of policymakers to analyse and study developments, as substantial resources have to be dedicated to solving these issues while, in numerous cases, leading to uncertain results.

Poor data quality introduces significant opacity and uncertainty, undermining the primary objective of data collection.

Systemic risk "builds up in the background before materialising"[8]: financial stability monitoring should prioritize frameworks that can proactively detect these risks. One of the cornerstones of effective risk detection is identifying specific anomalies in the data, such as pronounced risk concentrations or the emergence of large positions. This approach can be used in designing early warning systems, evaluating interconnectedness, assessing the implications of potential contagion risks, and evaluate possible tail events. For instance, daily monitoring of aggregate margin calls and their distribution during the March 2020 market turmoil has been key to identify risks[9] and build policy recommendations.[10]

While it is possible, from a technological and analytical viewpoint, to develop scalable and continuous monitoring framework,[11] the challenge lies in understanding the implications of data quality. Specifically, it becomes essential to distinguish whether an observed development is pertinent from a policy systemic risk viewpoint or merely a consequence of suboptimal data quality.

Policymakers are therefore faced with a challenge: that of **disentangling** relevant development from data quality issues. Given the size and complexity of the data, they need to tackle this problem in a scalable and – to the extent possible – automatable way across various datasets, counterparties, and markets. Moreover, it requires

[6] See ESRB (2020b)

[7] The European Securities and Markets Authority (ESMA) and the National Competent Authorities (NCAs) have taken several actions to improve the quality of EMIR data. These include the Data Quality Action Plan and the Data Quality Assessment Framework. See ESMA (2021).

[8] See Brunnermeier *et al.* (2012).

[9] See ESRB (2020a).

[10] See the Recommendation of European Systemic Risk Board of 25 May 2020 on liquidity risks arising from margin calls (ESRB/2020/6). Available at https://www.esrb.europa.eu/pub/pdf/recommendations/esrb.recommendation200608_on_liquidity_risks_arising_from_margin_calls~41c70f16b2.en.pdf

[11] See Abad (2016), Apicella et al (2022).

addressing data quality issues at any level of aggregation, thereby requiring approaching this problem in a way that allows to seamlessly shift across various levels of aggregation. For example, while the aggregate levels of a given quantity (e.g. notional amounts or initial margins) can be relatively stable in the aggregate, this may mask substantial concentration into one or few counterparties at the disaggregate level. Even upon identification, the presence of data quality issues does not allow to exclude the possibility that observed developments may arise from misreporting by an entity, such as inaccurate notional or margin amounts. This convergence of potential causes poses a challenge, which we refer to as "**anomaly intersection**".

In this paper, we introduce a framework to address this problem and apply it to granular, transaction-level data on derivatives collected under the European Market Infrastructure Regulation (EMIR) and available to the European Central Bank and the European Systemic Risk Board. This framework allows to make sense of observed anomalies in the data in two ways. First, it allows to set up rules at various levels of complexity to break down a potential signal across all its relevant dimensions. This facilitates the identification of which dimensions contribute to data quality issues or financial stability developments. Second, it provides a heuristic to reduce the likelihood of both false positives and false negatives by conditionally linking the detection of developments to the quality of the underlying data. More specifically, the probability that a development is relevant from a systemic risk viewpoint, increases in direct relation to the quality of the underlying data. Conversely, in those instances where a relevant financial development has been identified, its validation is contingent upon the framework indicating a superior data quality.

The framework allows to explore better the transaction-level datasets now available to policymaker: in the presence of dozens of dimensions, pinpointing the origin of an issue becomes a complex endeavor, requiring requires a robust system capable of efficiently navigating these dimensions, integrate expert judgment, and accounting for resource constraints. This paper proposes a first step to tackle this challenge.

2 EMIR data on derivatives and data quality

2.1 EMIR data – short overview

As a result of the turbulences caused by the global financial crisis, the 2009 G20 summit in Pittsburgh highlighted the need for greater transparency in the derivatives trading and put forward a set of measures intended to increase the stability of the international financial markets. Within the European Union, the objective was addressed through incentives to standardize derivatives contracts, introducing a mandate to centrally clear certain classes of derivatives via central counterparties (CCPs) and obligation to report them to trade repositories. The key legislation introduced to achieve those goals was the European Market Infrastructure Regulation (EMIR) (EU) No 648/2012.

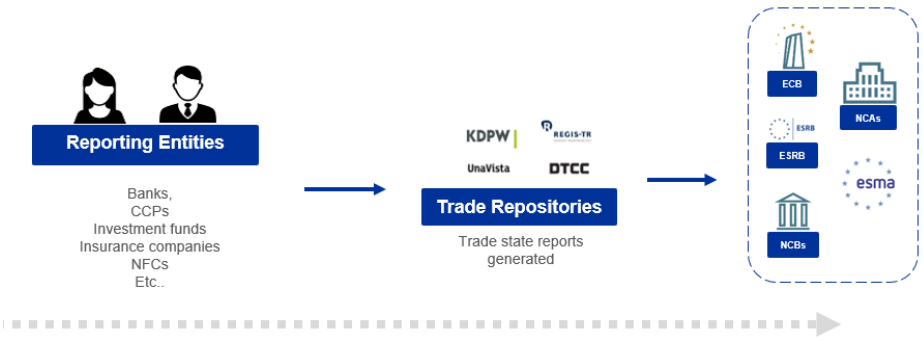
After the entry into force of the EMIR, EU competent authorities have been provided with an unprecedented amount of granular data. The data is reported daily

by all entities in the EU that are a counterparty to a derivative contract, both traded on exchanges as well as over the counter.

The decision to choose one or more of the TRs to which the trades will be reported is a free choice of the reporter and the deadline to report the transaction is the day after the transaction was executed, i.e. T+1. This implies that one reporter could submit the information of a particular trade either on the same day or the next date. In addition, the EMIR transaction reporting is a double-sided regime, meaning that every derivative trade has to be reported by both counterparties, as long as they are both resident in the EU. Thus, the two different reporting scenarios under EMIR from counterparties' perspective are (i) EU-EU and (ii) EU-non EU, while (iii) non EU-EU and (iv) non EU-non EU, are not reported under EMIR and may potentially fall under reporting rules of other jurisdictions. The typical transaction's reporting process is summarized in the Figure 1.

EMIR reporting process

Figure 1



Source: EMIR Regulation

Transaction reports received by the TRs consist of life-cycle events, that may represent, for instance, the conclusion, modification, valuation, and termination of a derivative. Owing to their volume, velocity, variety and veracity the EMIR data can be classified as "big data", which poses many challenges in using them.

2.2 EU data quality framework for EMIR

The entities reporting data under EMIR are responsible for delivering complete and correct information on concluded derivative contracts, in a timely manner. In practice, however, the data suffers from many inaccuracies, inconsistencies, or outright implausible values.

As a consequence, the authorities in the EU developed a comprehensive data quality framework to identify, exchange, prioritize and follow up on the issues found. The ESMA (European Securities and Markets Authority) assumes the leading, coordinating role in this process, directly supervising the TRs, and intermediating between authorities entitled to access EMIR data. The supervision of individual entities, however, lies within the remit of the relevant national supervisory agencies of EU member countries. This complex framework, together with predominantly cross-border nature of the derivatives trading in Europe, makes the task of dealing with reporting errors challenging.

An important tool in the supervisors' arsenal of methods to ensure data quality is the fact that the trades have to be reported by both counterparties to the trade, which allows the comparison of information transmitted in two separate reports, which, in turn, facilitates the identification of the incorrect reporting, and facilitates interpretation of potential anomalies discovered in the dataset. To support this endeavour, ESMA mandated the TRs to carry out a regular, weekly reconciliation exercise that identifies all the inconsistencies. The aggregated outcome is shared with the EU authorities.[12]

Nevertheless, despite the above efforts, the quality of EMIR data remains a significant problem. Given the size of the dataset, it's impracticable to expect that all data quality issues can be identified and addressed in a reactive manner. The importance of the regulatory reporting needs to be properly recognized by the reporting entities and should be further supported by clear and comprehensive reporting guidelines. Tackling the issues at the point of reporting data generation would be much more efficient than working on data quality on the receiving end and would enhance the quality of financial stability monitoring at the EU level.[13]

2.3 Classification of broad types of quality issues

Data quality is a multidimensional and complex concept. In the last decade, there has been a significant amount of work in the area of information and data quality management initiated by several research communities, ranging from techniques that assess information quality to build largescale data integration systems over heterogeneous data sources with different degrees of quality and trust. The development of established metrics to measure data quality is crucial to assess the significance of data-driven decisions.

EMIR requires market players to report an extensive set of characteristics for each derivative contract. Furthermore, financial companies (FCs), as well as non-financial companies (NFCs) above the so-called clearing threshold[14], are obliged to report daily valuation data corresponding to their open trades and positions, as well as any relevant updates to the value of collateral exchanged. This information is reported to TRs, and this process is often intermediated by third-party entities, e.g. exchanges, trading platforms, or reporting software providers.[15] The information received by the TRs is also used to create reports for the authorities. Along this reporting chain, data quality issues can emerge. Based on experience with data reported under EMIR, the following three main data quality categories can be identified and are discussed in this paper:

[12] As of 29 April 2024 (the go-live date of the so-called EMIR Refit) the TRs will also share the detailed reconciliation reports with reporting entities. See ESMA (2020a), section 6.2

[13] See ESRB (2022)

[14] As per EMIR Regulation (Article 4a and 10), NFCs and FCs are subject to the clearing obligation when exceeding predefined thresholds, see more details on ESMA's website: https://www.esma.europa.eu/policy-activities/post-trading/clearing-thresholds.

[15] It is important to note that EMIR Regulation (Article 9(1f)) permits the delegation of reporting to other entities (including entities not directly involved in the trade).

2.3.1 Data quality issues due to over-reporting

Over-reporting data quality issues represent the non-required but reported records (derivative contracts or their valuation updates) by the market players which in the case of EMIR data are challenged by the double-reporting regime and/or reporting delegation framework; alternatively, they could be triggered by duplicated records produced while being processed on the side of the TRs.

For example, EMIR applies to entities resident in the EU, thus the trades concluded with counterparties outside the EU are expected to appear only once in the dataset. Therefore, a transaction reported by entities from non-EU jurisdictions could fall into this case, similarly to derivatives which have not been terminated appropriately or have already matured but are still reported.

Some of the data quality issues may be also generated during the portability process, a process for transferring data from one TR to another.[16] The process was widely employed in the context of Brexit, but it is also frequently triggered by reporting entities on voluntary basis. Errors or inaccuracies in the transfer of information between the TRs can lead to duplicated transactions in the final dataset.

2.3.2 Data quality issues due to under-reporting

Under-reporting data quality issues represent the required but not reported records (derivative contracts or their valuation updates) by the market players. Alternatively, they could be existing records gone missing while processing the data on the side of the TRs.

A straightforward example could be a reporting entity that does not comply with EMIR reporting obligation and does not report its derivative contracts to the TR. Varying input formats required by the TRs could also be a potential data quality issue as the data needs to be transformed into the XML-based ISO20022 message, which the TRs are obliged to provide to authorities.[17] In this conversion process, due to lack of standardisation of the information submitted or due to the incorrect mapping implemented by TRs, some records might be rejected from the final pool of transactions available for analysis, as they do not conform to the schema of the message to be transmitted to authorities. Similarly, transactions which fail the ESMA validation[18] rules could be rejected by the TRs and may never reach the authorities.

Related to the abovementioned portability process, errors or inaccuracies in the transfer of information between the TRs can also lead to missing transactions in the final dataset.

Moreover, some submissions being reported late could be misinterpreted or not captured on a timely basis, and in turn they could bias the reconciliation of derivatives

[16] See ESMA (2017)

[17] This lack of input standardisation is expected to be significantly mitigated by the upcoming change of EMIR technical standards, so-called EMIR Refit, see: https://www.esma.europa.eu/sites/default/files/library/esma71-99-1490_press_release_emir_refit_final_report.pdf

[18] See https://www.esma.europa.eu/policy-rules/post-trading/trade-reporting.

contracts. The reconciliation of derivatives contracts, described in section 2.2, is a relevant component for the data quality assessment, e.g. in the context of estimation of non-reported trades subject to double-sided reporting.

One exemption to keep in mind related to EMIR is that the regulation does not impose the reporting obligation on natural persons, hence for trades carried out by private individuals only the leg reported by the legal entity will be visible in the final dataset.

2.3.3 Data quality issues due to misreporting

Misreporting data quality issues arise in the required and reported records (derivative contracts or their valuation updates) containing erroneous information.

The erroneous information may come from the internal system of the reporting entities or be introduced in the process of the transformation of the details of the trades to TRs. It may consist in a variety of issues, including simple typos, incorrect categorisation or numerical values, as well as inaccurate interpretation of reporting guidelines[19]. The failure of CCPs, clearing members or more generally market players not coordinating on trade ID, counterparty ID or on position- versus transaction-level reporting[20] could also lead to the impossibility to reconcile trades subject to double-sided reporting, impacting the final data quality assessment and the overall analysis.

As in the above cases, the accuracy of the mapping from TRs' proprietary input formats to the XML message and the ESMA validation rules also plays a pivotal role in the accuracy and presence of the records in the final dataset.

Some of the data quality issues categorised above may be also caused by the complexities associated with the full life-cycle of a set of contracts. These may also include the post-trade processing, such as clearing, netting or compression of derivative contracts[21]. These complexities could lead to difficulties in correctly representing those events in the reporting template, and consequently to data quality issues listed above.

Distinguishing between data quality issues and genuine developments (e.g. market shifts) in EMIR is not straightforward. For example, a sudden increase in the volumes traded and reported submissions of a specific entity may have various reasons – it may represent a change in the trading behaviour, affected by the volatility of the market. It could, however, also stem from reporting errors. The anomaly-detection algorithms, controlling the changes of the outstanding amount over two specific dates, may not be in a position to tell apart those two scenarios, and may consequently generate misleading signals to competent authorities monitoring the

[19] For regulatory technical standards and implementing technical standards, see https://www.esma.europa.eu/policy-rules/post-trading/trade-reporting.

[20] Position level reporting means net positions resulting from a set of contracts representing fungible products, rather than per individual trade. See TR question 17 in "ESMA Question and Answers on EMIR Implementation": https://www.esma.europa.eu/sites/default/files/library/esma70-1861941480-52_qa_on_emir_implementation.pdf

[21] See ESMA (2020)

developments in the derivatives markets. This challenge will be further discussed in chapter 5 .

2.4 Challenges in data quality assurance in large-scale financial datasets

In the traditional data warehouse environment, comprehensive and manual data quality assessment and reporting was possible (if not ideal). Ensuring data quality could be thought of as a four-step approach: i. selecting data quality dimensions; ii. designing an assessment plan; iii. assessing the plan against the selected dimensions; iv. acting on results. The elements which are traditionally included in such an assessment plan are the following:[22]

- **Validity:** the data is adherent with precision to a given real-world phenomenon
- **Reliability:** the data is defined, measured and collected in the same way all the time with a high degree of trust and reputation
- **Completeness:** the data contains all the information with pertinence at the set, record and element levels
- **Accuracy:** the data is detailed and correct and the units of measure are clear and valid
- **Timeliness:** the data is available on time
- **Integrity:** the data is internally consistent and not biased
- **Uniqueness:** the data does not contain the same information more than once.

However, in Big Data projects the scale of data makes the above process challenging. Thus, in many cases, the data quality measurements can at best be approximations, i.e. need to be described in probability and confidence intervals, and not in terms of absolute values. The challenges posed are mainly driven by the intrinsic features of large-scale financial datasets:[23]

- **Volume:** the large-scale amount of data poses analytical challenges as it requires advanced handling techniques (e.g. parallelisation, partitioning and clustering) within a reasonable overhead on time and resources (storage, compute, human effort, etc.)
- **Velocity:** the high-speed of the reporting, collection, processing, visualisation and transformation of data into targeted insights poses timely challenges as by the time data quality assessment is completed, the output might be outdated, therefore it requires advanced processing techniques (e.g. sampling)
- **Variety:** for efficiency purposes, the data might include several data types (structured, semi-structured, and unstructured) coming in from different sources, therefore it requires advanced modelling techniques (e.g. structured metrics)
- **Veracity:** the amount of bias, noise and abnormality might hinder the accuracy and reliability of the dataset, making it difficult to add value created by

[22] See Loschin (2010), Chapter "The Organizational Data Quality Program"

[23] See Du (2018), Chapter "Overview of Big Data and Hive"

identifying new patterns and fostering the decision-making process, therefore it requires advanced decision-making techniques (e.g. identification and classification)

- **Visualization**: visual loss due to noise of the excessive information available.

Differentiating the data quality dimensions is the key for matching potential issues against a business need and prioritizing which dimensions to assess and in which order becomes the problem to solve for large-scale datasets. In order to handle big datasets, another challenge is choosing which among the following data reduction strategies to apply:

- **Sampling:** every dataset can be viewed as a sample; the latter is featured by a probability value which can return the fraction of data with representative properties as a result
- **Filtering:** filtering for a specific dimension (e.g. timing) which meets specific conditions is another technique to query large datasets
- **Aggregation:** the dataset grouped by records falling within predefined bins into subsets.

The expression "Big Data" does not simply refer to its vast amount of information but it intrinsically recalls the technology, processes and techniques employed to store, manipulate and share the information on a large scale. On the basis of the difficulties posed by the size of the data and the intrinsic features of the underlying transactions, careful design is necessary in the systems used for data collection and analysis to ensure that the output actually produces some insightful content correctly interpreted.

3 Methods

3.1 Modelling framework

The purpose of the Automated Data Quality tool is to identify and classify the developments in numerical measures of granular, multi-dimensional datasets of financial information. Let's assume that we have a collection of N_t observations, where each observation reflects details of an individual financial phenomenon, e.g. transaction, instrument, or lifecycle event. In the application to EMIR, each observation will represent the aggregate position of a counterparty on a specific type of derivatives. Furthermore, this information is available for two points in time, described as reference periods, denoted t and $t-1$.

The dataset can be then described as the following matrix:[24]

$$X_t = [c \quad d^1 \quad \cdots \quad d^U \quad m]$$

[24] Please note that a wider dataset with additional columns or higher granularity can be easily converted to such representation by a series of SELECT, GROUP BY, and WHERE operations. The ADQ tool carries out such pre-processing of the dataset within the dimension and entity-level steps, described in section 4.1.

where:

c – a column (vector) identifying the entity involved in the observation, e.g. one of the counterparties; the values of c are alphanumeric strings, uniquely identifying the entity; the value cannot be empty. We denote the unique values in column c as ID^x, hence $c_i \in \{ID^1, ID^2, \dots, ID^X\}$;

d^u – a column (vector) representing dimension u of the U categorical dimensions describing the characteristics of the observation, $u \in \{1, \dots, U\}$; each element of the column can take a value from a predefined list $d_i^u \in \{v_1^u, v_2^u, \dots, v_{Z_u}^u, NULL\}$, where $NULL$ indicates that the given dimension is not available or not relevant for the observation in question;

m – a column (vector) representing a numerical measure[25] describing the observation, e.g. market value of the contract; it is assumed that a value of NULL is equivalent to 0 for this column.

We calculate the total delta Δ as the difference of the sum of values of measure m, in two datasets pertaining to reference periods t and $t-1$:

$$\Delta = \sum_{i=1}^{N_t} m_{i,t} - \sum_{j=1}^{N_{t-1}} m_{j,t-1}$$

However, given that the changes in different observations can have opposing direction, we also define total absolute delta Δ^{abs}, reflecting the sum of absolute differences between observations characterized by the same dimensions $[c, d^1, d^2, \dots, d^U]$. When determining Δ^{abs}, we need to ensure that we calculate the individual Δ's between the measures referring to the same dimensions. For this purpose, each matrix is supplemented with missing sets of dimensions associated with measure equal to 0. As a consequence, both matrices will contain N' observations, where $N' \geq N_t$ and $N' \geq N_{t-1}$.

$$\Delta^{abs} = \sum_{i=1}^{N'} \Delta_i^{abs} = \sum_{i=1}^{N'} |m_{i,t} - m_{i,t-1}|$$

The goal of the tool is to identify a set of K row vectors $r_k, k \in \{1, \dots, K\}$, each consisting of a set of conditions[26] taking one of the following forms:

- $c = ID^x$
- $c \neq ID^x$
- $d^u = v_z^u$
- $d^u \neq v_z^u$

Each r_k identifies a set of conditions, which can be mapped to a set of observations $[ID^i \quad v_i^1 \quad \cdots \quad v_i^U]$ contained in the matrices X_t and X_{t-1}. In other words, vectors r_k partition the matrices X_t and X_{t-1} in K pairwise disjoint sets. The vectors r_k have to be mutually exclusive, i.e. none of the observations $[ID^i \quad v_i^1 \quad \cdots \quad v_i^U]$ should be mapped at the same time to two different vectors r' and r''. Consequently, each vector r_k can be assigned a portion of Δ^{abs},

[25] The measure can take either positive or negative values.

[26] Not all dimensions have to be included in the r_k vector, and conversely one dimension may be subject to multiple conditions. See example below.

corresponding to the respective set of observations. If we denote the part of Δ^{abs} corresponding to the vector r_k as $f(r_k)$, we can write:

$$f(r_k) = \Delta_k^{abs}$$

Consequently, we are looking for an algorithm satisfying the following condition:

$$\Delta^{abs} = \sum_{k=1}^{K} f(r_k)$$

There exist multiple partitions $[r_1, r_2, \dots, r_k]$ of the matrices X_t and X_{t-1}. In the extreme, a collection of vectors representing each possible permutations of allowable values of $c, d^1, d^2, \dots, d^U$ would explain the entirety of Δ. On the other hand, an empty vector would do the same.[27] Both solutions are not satisfactory from explanatory point of view. The purpose of the algorithm described in the following section is achieving certain explanatory power, i.e. find a partition that allows us to attribute anomalies in the dataset to a limited number of observation subsets, each of them described by a set of conditions imposed on the dataset dimensions.

3.2 Algorithm

In order to select the solution with optimal explanatory power, and able to tackle the enormous size of the granular regulatory datasets, the ADQ employs binary trees (see Nasiriany (2019), Chapter 7) to select the relevant dimensions d^u and their corresponding values (i.e. a partition $[r_1, r_2, \dots, r_k]$), which contribute to the developments in the dataset.

Nevertheless, this approach is not sufficiently performant for the entity identifiers, as their number can easily reach hundreds of thousands unique values in datasets available to the ECB and ESRB.[28] Therefore, the approach consists of two modules, which can be triggered independently, or consecutively, depending on need:

- entity-level analysis module,
- dimensions' analysis module.

3.2.1 Entity-level analysis module

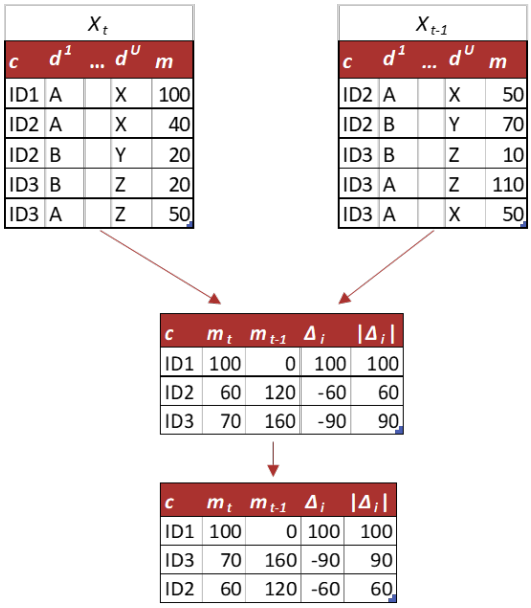
Given that the number of entity columns is restricted to one, the selection of main entities explaining the developments in the dataset simplifies to a simple JOIN / GROUP BY operation, as illustrated in Figure 2.

[27] The empty vector should be interpreted as lack of filtering conditions imposed on the dataset, hence all allowable combinations would be covered by an empty vector r_k.

[28] Not taking into consideration the performance constraints, the problem could be simplified by treating the c column as one of the dimension column d^u.

Entity-level analysis – numerical example

Figure 2



Source: Own calculations

Notes: The chart presents a numerical example of the JOIN / GROUP BY operations performed on the dataset to identify entities contributing most to the change in the analysed measure.

The outcome of those operations clearly indicates the entities contributing most to the change in measure in question.

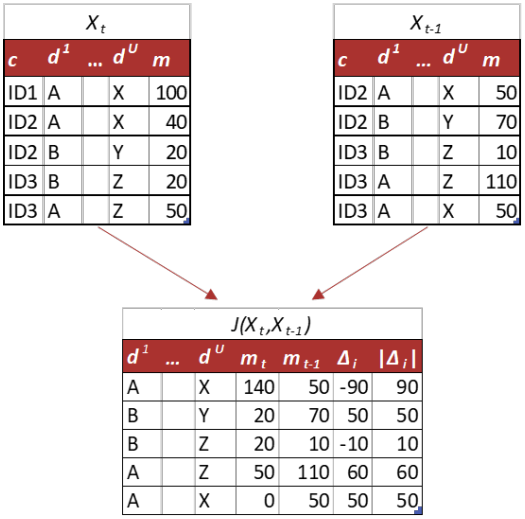
3.2.2 Dimensions' analysis module

The analysis of dimensions is more complex. Following on the example in section 3.2.1 a set of JOIN / GROUP BY operations allows us to calculate Δ per combination of dimension values (we denote the outcome of this operation as $J(X_t, X_{t-1})$:[29]

[29] Please note that the sum of absolute Δ varies between the entity-level and dimensions' analysis modules. This is caused by the fact that for each module the absolute value is calculated for sub-aggregates on different level of aggregation.

Dimensions' analysis, initial grouping – numerical example

Figure 3



X_t

c	d^1	...	d^U	m
ID1	A		X	100
ID2	A		X	40
ID2	B		Y	20
ID3	B		Z	20
ID3	A		Z	50

X_{t-1}

c	d^1	...	d^U	m
ID2	A		X	50
ID2	B		Y	70
ID3	B		Z	10
ID3	A		Z	110
ID3	A		X	50

$J(X_t, X_{t-1})$

d^1	...	d^U	m_t	m_{t-1}	Δ_i	$\lvert\Delta_i\rvert$
A		X	140	50	-90	90
B		Y	20	70	50	50
B		Z	20	10	-10	10
A		Z	50	110	60	60
A		X	0	50	50	50

Source: Own calculations

Notes: The figure presents an example of JOIN / GROUP BY operations, performed on an example dataset, in order to calculate the absolute changes characterising different dimension sets.

In this case, however, we have to count with multiple dimensions, and potentially hundreds of thousands of combinations of dimension values. Individual absolute deltas cannot give us any practical insight into which dimension contributes most to the change in the dataset.

To tackle this problem, we employ binary decision trees. We construct a tree where each decision reflects the dimension and corresponding value that best characterize the change in the investigated measure. At each node the tree is split into sub-trees representing subsets of $J(X_t, X_{t-1})$. In order to determine the optimal split, the tool measures the *Gini impurity index*:[30,31]

$$\text{gini} = 1 - p^2 - (1-p)^2 = 2p(1-p)$$

where $p = \frac{|\Delta_i|}{\sum |\Delta_i|}$, i.e. the share of the absolute delta corresponding to particular combination of dimensions in the total absolute Δ.[32] The split by dimension with minimum Gini impurity index value provides the highest contribution to the explanation of the development of the measure under examination.

To illustrate the procedure, let's assume that we consider only two dimensions d: "contract type" and "asset class".[33]

[30] See Nasiriany (2019), p. 166

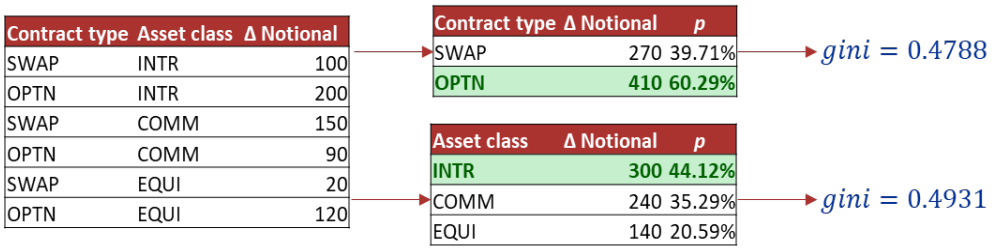
[31] The tool can also apply the entropy measure in place of the Gini impurity index.

[32] It is important to note that this method requires that the absolute value is calculated on the most granular level, and then summed up in the following steps. Otherwise, the results calculated on different nodes of the tree would not be additive and depending on the path taken by the algorithm we would arrive at different dataset splits.

[33] The following abbreviations apply: SWAP = swap, OPTN = option, INTR = interest rate, COMM = commodity, EQUI = equity.

Calculation of Gini impurity index – numerical example

Figure 4



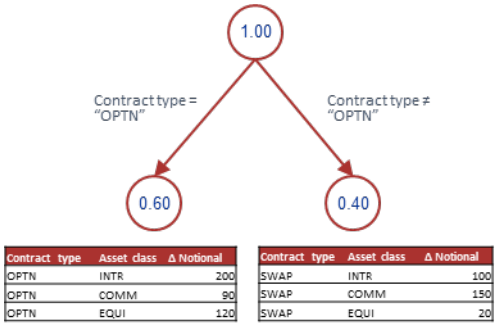
Source: Own calculations

Notes: The figure presents a calculation of the Gini impurity index for two possible groupings of the example dataset

As the Gini impurity index is lower for the "contract type" dimension, the tree is first split into two branches according to the criterion $\text{contract type} = \text{"OPTN"}$ and $\text{contract type} \neq \text{"OPTN"}$. For each sub-tree the weight of the tree is calculated, reflecting the share of the total absolute Δ explained by the sub-tree. The procedure is applied recursively, according to some pre-determined stopping criteria.[34]

Construction of the decision tree, top node – numerical example

Figure 5



Source: Own calculations.

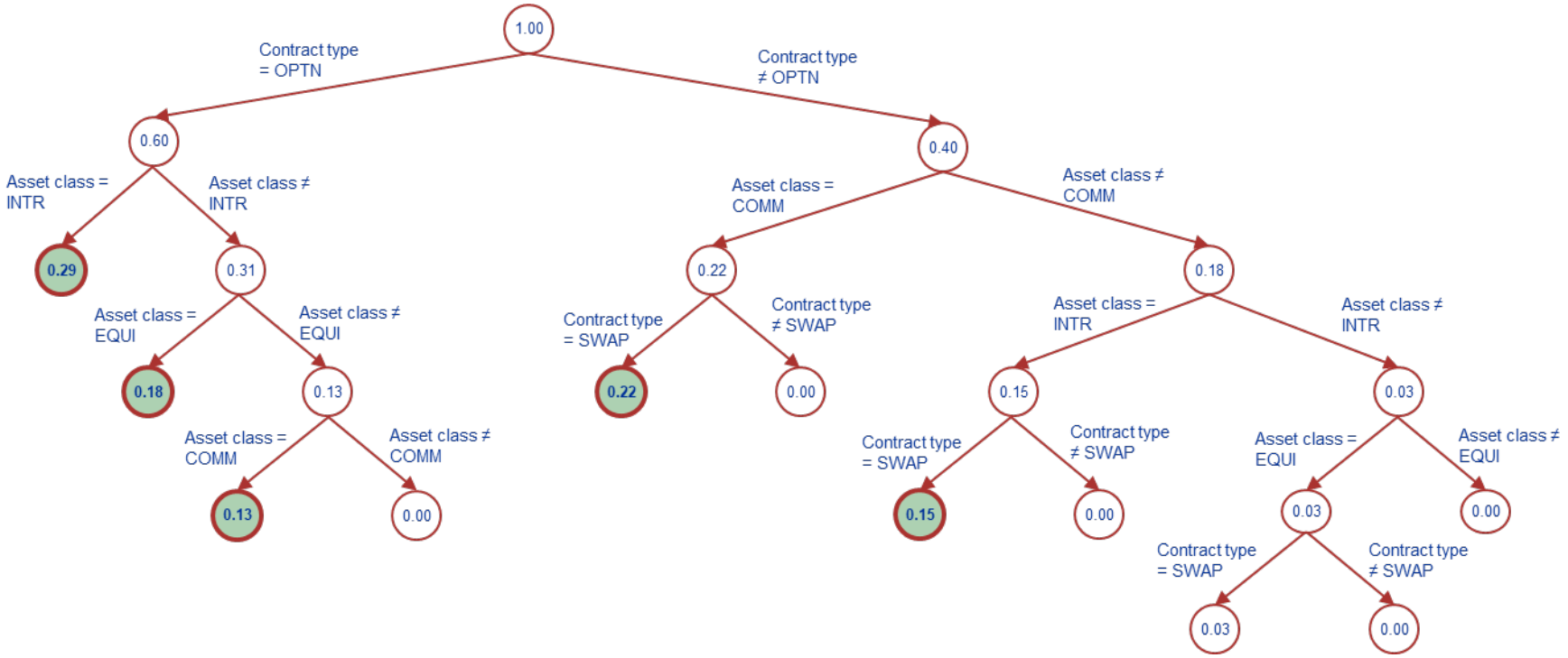
Notes: The figure presents a stylised example of a subtree, split by the $\text{contract type} =/\neq \text{OPTN}$. The numerical values inside the circles denote the share of the delta explained by the conditions on the path leading to the respective node. The tables below the child nodes represent the subsets of the dataset, according to the split criterion.

In the simple example above, the algorithm leads to the construction of the following tree, where the leaves with ultimate weight exceeding a predetermined threshold (in this case 0.1) are highlighted in green:

[34] The ADQ applies two customisable stopping criteria: maximum depth and minimum threshold change in impurity measure.

Construction of the full decision tree – numerical example

Figure 6



Source: Own calculations.

Notes: Notes: The figure presents a stylised example of a tree built on the basis of the dataset presented in the Figure 4. Each branch is split according to a criterion that minimises the Gini impurity index, until predetermined stopping criteria are fulfilled. The numerical values inside the circles denote the share of the delta explained by the conditions on the path leading to the respective node.

The weight of the leaf is also an indication of how much the conditions along the path leading to it contributed to the change in the total absolute Δ. From the diagram above we can see that the change in the measure over time was driven mainly by interest rate options (29%, INTR and OPTN) and commodity swaps (22%, COMM and SWAP).

Each path leading to a leaf in the constructed tree represents a vector r_k, as described in section 3.1, while the number in the circle is the share of the absolute delta, corresponding to this particular vector.[35]

Obviously, in this simple example, the algorithm does not constitute a material advantage over visual inspection of the dataset. However, in a scenario with dozens of dimensions and millions of transactions, any manual or semi-manual approach clearly falls short.

3.3 Application to double-sided reporting reconciliation

As stated in section 2.1, some collections of granular financial data foresee the reporting of a particular financial phenomenon (e.g. derivative transaction) by both counterparties linked to the transactions. This type of collection is commonly described as "double-sided reporting", and data reported under EMIR fall into this category.[36] Under the assumption of correct reporting, it can be expected that information referring to the characteristics of the trade is consistent between the sets of data reported by two involved counterparties. Similarly, the quantitative measures describing the contract should coincide if the measurement is made at the same point of time.

The double-sided reporting offers a unique possibility to benchmark the quality of the information reported by the counterparties. If the information reported differs significantly between the two reporting entities, it can be concluded that one of them (if not both) reports incorrect information, or does not report some data at all.

In the following analysis we will focus on the discrepancies in the reporting of the quantitative measures reported by the counterparties. For the purpose of assessing the quality of the reported information and understanding the underlying reasons, it is important to determine: (i) what are the pairs of entities that exhibit largest discrepancies, and (ii) if there are any specific characteristics of the observations underlying those differences that could give additional insight into the reasons for the discrepancies.

The procedure developed in section 3.1 can be easily adjusted to the case of double-sided reporting. Let's assume that we have a dataset of financial information in the following form:

$$X = [c^R \quad c^O \quad d^1 \quad \dots \quad d^U \quad m]$$

The notation from section 3.1 applies accordingly with the following additions:

[35] To be precise, vector r_k, includes also the counterparty identifier c, identified in the entity-level analysis module (see section 3.2.1). See discussion in section 4.1 on how the two models interact in practice.

[36] This does not apply if one of the counterparties is resident outside of the EU, or is a private individual.

c^R – identification of the counterparty that reported the observation in question, also "reporting counterparty"

c^O – identification of the other counterparty that was involved in the observation, also "other counterparty"

The dataset X is double-sided, if and only if:

$$\forall i \exists j (c_i^R = c_j^O \wedge c_j^R = c_i^O)$$

In other words, for each observation there exists a corresponding one, with the same pair of counterparties, but in reverse. A tuple $\{i, j\}$, satisfying the condition $(c_i^R = c_j^O \wedge c_j^R = c_i^O)$ will be further described as a "paired position".

The dataset X can be split into two disjoint sets X' and X'', such that elements of each paired position are separated, namely:[37]

$$\forall_{i,j} (c_i^R = c_j^O \wedge c_j^R = c_i^O) \rightarrow (X_i \in X' \wedge X_j \in X'') \vee (X_i \in X'' \wedge X_j \in X')$$

where c^R and c^O columns are replaced by a new identifier c^{pair}, which is a concatenation of the c^R and c^O identifiers, ordered alphabetically.[38] In this way the c^{pair} becomes a key, linking paired positions segregated into X' and X''. The split criteria can be arbitrary, although, obviously, it is reasonable to assume the criteria following certain business logic. For instance, in case of a dataset that contains transactions between Central Clearing Counterparties (CCPs) and Clearing Members (CMs), it is reasonable to split X along the criterion $c^R \in CCPs$ / $c^R \in CMs$. In other cases, an artificial splitting criterion may be needed, for example an alphabetical ordering of entities' IDs.

Consequently, we arrive at two datasets:

$$X' = [c^{\text{pair}\prime} \quad d^{1\prime} \quad \dots \quad d^{U\prime} \quad m']$$
$$X'' = [c^{\text{pair}\prime\prime} \quad d^{1\prime\prime} \quad \dots \quad d^{U\prime\prime} \quad m'']$$

In this way the above problem reduces to the one described in section 3.1, with the datasets X' and X'' corresponding to X_t and X_{t-1}, respectively. Applying the algorithm described in section 3.2 results in identifying the largest differences in the information reported by counterparties, as well as explaining any patterns in characteristics of the trades, for which the differences occur.

The above reasoning can be also applied to reconciliation of other types of information, which is represented by two disjoint subsets of data referring to a particular reference period.[39] For brevity, we denote the algorithm described in section 3.2 as "time-series analysis", and the one characterised in section 3.3 as "intraday analysis".

[38] For example, if $c_i^R =$ "$ABC1$" and $c_i^O =$ "$XYZ2$", then $c_i^{\text{pair}} =$ "$ABC1; XYZ2$"

[39] One example could be long and short positions of a CCP in specific products – the CCP by construction should have no net exposures, i.e. the absolute value of the short position should be equal to the long position.

4 Application of the ADQ method to EMIR data

4.1 ADQ Process

The ADQ method finds an application in a large-scale dataset such as EMIR. It allows to identify timely **data quality issues** and **developments of the derivatives market** overcoming the challenges posed by the size of the dataset. The application leverages on the EMIR IT system that is in place at the ECB since 2017. The ECB implementation relies on a Hadoop infrastructure and allows data consumers at the ECB to perform certain analytical activities that previously took hours or days, in a matter of minutes. The ADQ process applied to EMIR data is integrated in a set of automated daily activities, such as processing, enrichment,[40] and data quality management, carried out at the ECB to ensure the timely provision of EMIR data to users.

The method is applied to the trade state reports of EMIR[41] that include the outstanding trades on a given date (i.e. reference period). The input to the process is provided by a set of parameters and by monitoring information resulting from the daily processing of the EMIR data. The measures, the dimensions, the concentration threshold, the input dataset, and the type of the analysis (time-series vs. intraday) are defined in the set of parameters that is provided to the process in the form of a JSON file. Varying the set of parameters allows running different jobs to analyse the dataset from different points of view. The monitoring information, in turn, allows to assess the completeness of the data reported by the trade repositories (TRs) on a given reference period. In case of uncomplete data submitted by a TR, all data from this TR are removed from the dataset that will be analysed.

Once these preliminary processes are carried out, we move to the analysis of the data according to the two modules: entity-level analysis and dimensions' analysis. These are executed sequentially and in two parallel workflows:

1. *Entity-level analysis followed by dimensions' analysis*: first the main entities contributing the most to the changes in the quantitative measure are selected and then the dimensions' analysis is applied to these entities to determine the driving factors of the changes observed.
2. *Dimensions' analysis followed by entity-level analysis*: first the combinations of dimensions with the largest explanatory power of the changes in the quantitative measure are selected and then the module of the entity-level analysis is applied to determine the main players responsible for the changes observed.

Once both workflows are concluded, the results are summarized in an HTML report circulated via e-mail to the stakeholders of the EMIR dataset. Such report

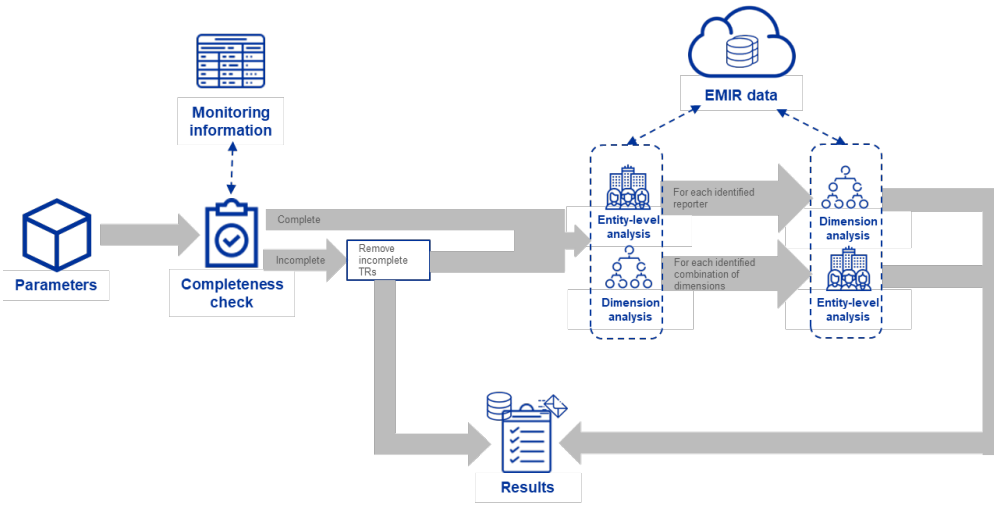
[40] The enrichment is a process, where the dataset is complemented with information from other reference datasets. At the ECB the data collected under EMIR is enriched with supplementary information on entities, benchmarks and underlying instruments from several internal and external sources.

[41] The TRs provide the authorities with two main types of reports:

- trade state: information on all derivative contracts outstanding on a given reference date;
- trade activity: information on new trades and lifecycle events affecting existing trades reported within a particular reference date.

provides information on the set of parameters applied, the main entities and dimensions that drive the changes observed in the quantitative measure resulting from the two workflows. In addition, the results of the ADQ algorithm are stored in a database that allows for feeding other processes and systems, for instance graphical tools to visualise the results in charts and dashboards.

ADQ architecture Figure 7



Source: Own work

4.2 What do we measure?

The EMIR data include several quantitative measures that can be analysed through the ADQ method, such as the notional, the contract value and the initial margin received defined according to the EMIR Regulatory Technical Standards[42] as follows.

- **Notional:** The reference amount from which contractual payments are determined. In case of partial terminations, amortisations and in case of contracts where the notional, due to the characteristics of the contract, varies over time, it shall reflect the remaining notional after the change took place.
- **Value of contract value**: Mark to market valuation of the contract, or mark to model valuation where applicable under Article 11(2) of Regulation (EU) No 648/2012. The CCP's valuation are to be used for a cleared trade.
- **Initial margin received:** Value of the initial margin received by the reporting counterparty from the other counterparty.

A set of parameters for each measure is created and passed to the ADQ process together with all the other relevant pieces of information, for instance the source

[42] Commission Delegated Regulation (EU) No 148/2013 of 19 December 2012 supplementing Regulation (EU) No 648/2012 of the European Parliament and of the Council on OTC derivatives, central counterparties and trade repositories with regard to regulatory technical standards on the minimum details of the data to be reported to trade repositories: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02013R0148-20171101&from=EN

dataset or the output database where the results will be stored. This results in different instances of the process that run independently for each measure.

Part of the initial set of parameters are the pieces of information to configure the two ADQ modules, namely the entity-level and the dimensions' analysis. The former requires as input information the identifier of the entity, for instance the **reporting counterparty** of a trade; the other module is specified by a set of dimensions. We identified 8 dimensions of the trade that are significant to explain the changes observed in EMIR data: the **asset class**, the **contract type**, the **currency** of the notional, the **clearing flag**, the **intragroup flag**, the **execution venue**, the **type of observation** reported in the trade activity report,[43] and the **trade repository** that submits the trade to authorities.

Another significant element characterising each instance of the ADQ process is the type of analysis. We differentiate between **time series** and **intraday** analysis:

1- *Time series analysis*: the trade state table at T is compared with the one at T-1. This type of analysis allows detecting data quality issues such as implausible values of notional and contract values for several reporting entities. In addition, the analysis over time of the data provides insights to market developments. This is the case of the increase of initial margins occurred with the outbreak of the pandemic (March 2020) or the movement of initial margins due to the developments in the European gas market in October 2021.

2- *Intraday analysis*: the information reported at T by the two legs of trades is compared. The intraday analysis is applied to two categories of reporting entities: CCPs and clearing members (CMs). The aim is to measure the discrepancy between the notional reported by pairs of CCPs and CMs. Therefore, the application of the method requires the aggregation of raw data computing the total notional by each pair CCP-CM and CM-CCP. This allows to detect issues of under-/over-reporting by one of the two sides for specific trades defined by the dimensions.

In the current set-up at the ECB, multiple workflows are triggered automatically every morning covering the above types of analysis with the run-time amounting to 5-10 minutes per workflow. The outcome of those workflows is shared with the group of EMIR operators, who then act on the findings. Conditional on the type of the issue, the matter may be further investigated, reported to an adequate authority (e.g. ESMA), raised to the attention of the respective trade repository and/or shared with internal users of the dataset.

4.3 Extension of the work

The flexibility and customisability of the process are ensured by its full parametrisation. This is not limited to the selection of the quantitative measures and dimensions to be explored but it also includes the possibility of running preliminary transformation to raw data that will serve as input data to be analysed. Therefore, the

[43] I.e. transaction- or position-level record, see also footnote 20.

method can be easily extended both in terms of instances within EMIR data perimeter and beyond to other datasets.

Considering further applications to EMIR, further workstreams could be implemented, such as

- additional module for the processing of data quality issues detected for transmission to ESMA as authority in charge of the EMIR data quality management.
- Application to trade state reports and trade activity reports to analyse the consistency between the two kinds of reports.
- Correcting for the potential temporal misalignment of reporting, e.g. when entities report the corresponding information on different days.
- Time series analysis applied with a larger lag, i.e. compare data at T with the data at T-30.
- Building complex customisable pipelines from ADQ modules, e.g. applying sequentially the entity-level analysis on reporting entities followed by the entity-level analysis on the other counterparties to the trades for the 3 main reporting entities and then concluding with the dimensions' analysis.

Regarding the extension to other datasets, the work started already and will be further complemented to apply the method to SFTR data both implementing the time series analysis and the intraday one to CCPs and CMs.

5 Disentangling anomalies

Detecting anomalies in the financial system through the analysis of a broad set of indicators represents the foundation of systemic risk monitoring. This set of indicators typically include the build-up of large exposures, concentration and interconnectedness, or the identification of specific exposures that are particularly sensitive under certain scenarios (e.g. to repricing and margin calls). Once detected, these anomalies could signal relevant financial stability developments and inform policymakers' actions.

However, in case of low quality of the data reported by market participants, these anomalies may simply reflect inaccurate information, rather than a development relevant from a financial stability viewpoint. In turn, this generates uncertainty in interpreting analytical results, which impairs monitoring capabilities, potentially leading to wrong conclusions. Additionally, uncertainty can lead to rely less on the data, thus spend less effort on its analysis, which can further worsen their quality as more issues are undetected. Avoiding this self-fulfilling spiral should therefore be a strategic objective of regulators.

From a research standpoint, low data quality can be often dealt with by narrowing the analysis by selectively restricting samples (e.g. to remove implausible observations), correcting outliers, or making specific assumptions. From a policy perspective, however, the downstream impact of both low data quality and the potential assumptions to deal with it can be significant if not carefully considered. In

fact, working in the presence of uncertainty creates both analytical and operative issues.

First, from an analytical perspective, the low data quality reduces the reliability of the results opens to potential false positives (e.g. when a substantially high value for an indicator measuring concentration is due to erroneous data) or false negatives (e.g. a low value of a relevant bilateral exposure due to missing or erroneous data). In this case, policymakers need to embark in a time-consuming case-by-case inspection to understand the potential root causes and gauge the impact of low-quality data. Moreover, they may have to judge whether the impact of low data quality is material or not, adding further assumptions to the analysis.

Second, from an operative perspective, low data quality makes financial stability monitoring substantially more challenging when it needs to be performed "at scale" and only partly automated: working case by case is not operationally feasible in the presence of very large datasets, reported with high frequency, e.g. daily. The rationale to scale up analytical systemic risk monitoring lies not only in the size of newly available data, but also on the evolving nature of risks in an increasingly complex, interconnected, and adaptive financial system. Analytical scaling also shows an intrinsic dimensionality problem: the number of potential indicators and their levels of aggregation can become easily extremely large.

In this section, we outline an approach to use the framework illustrated in this paper to mitigate this problem. The main intuition underpinning this application is straightforward. By *disentangling* between the two main sources of anomaly in the data, we can reduce the odds of encountering both false positives and false negatives. If the data is of high quality, the probability that an anomaly is a significant financial stability development is higher, whereas if the data is of low quality, this probability decreases. Leveraging insights gained from the ADQ framework, policymakers can discern genuine financial stability signals with less uncertainty.

The key feature of the ADQ framework is its capability to pinpoint the primary contributors to data quality issues by progressively breaking down along the relevant dimensions. This allows policymakers to make informed methodological decisions when interpreting a financial stability signal in the presence of suboptimal data quality.

Utilizing the ADQ framework can reduce uncertainties related to analytical outcomes. The framework offers at least two ways to achieve this: upstream and downstream.

1) **Upstream**. The first way is to start from the anomaly, as detected by the ADQ tool, and then analyse the quality of the underlying data. Once an anomaly is detected via the ADQ framework in the time-series mode, the ADQ can be applied in intraday mode on the two dates: if the ADQ in intraday mode does not lead to a data quality issue on both dates, then the anomaly detected in time-series mode is less likely due to a data quality issue. On the contrary, if the ADQ run in intraday mode returns a data quality issue, then the anomaly can be attributed to low data quality, according to which day it has appeared.
2) **Downstream**. The second way to reduce uncertainty via the ADQ is to start from a data quality issue and understand which analyses this may impact downstream: any anomaly detected which uses observations for which it is known that a data quality issue is present will have a higher likelihood to be due to a data quality issue.

It is important to remark that, while this use of the ADQ framework can be helpful to facilitate policymakers' work, it can never substitute the value of having high quality data. In the following two examples, we are going to illustrate how this approach can be used.

Example 1: margin calls. The first example uses the ADQ framework to disentangle the anomaly upstream. Let us imagine we observe a substantial increase in the margins reported by a Central Counterparty during a crisis period, detected by running the ADQ in time-series mode. While one may have anecdotal knowledge of potential margin calls, it is still unsure whether the margin call is in the order of magnitude signaled by the CCP and who are the clearing members, products, and clients affected by these margin calls. To this end, the first step of the disentangling procedure would be to run an ADQ procedure on the delta between two different dates to understand the relevant dimensions (e.g. the clearing members). Let us now imagine that the margin call is explained by a substantial fraction (e.g. more than 50% of the total margin increase) by one individual clearing member and the policymaker is unsure whether this is a relevant financial stability signal or it is due to a problem in the data reported by the CCP. If running a further ADQ process in intraday mode on the margin reported from the clearing members' perspective shows no data quality issue, then the anomaly is likely relevant from a financial stability viewpoint. If, on the contrary, a discrepancy between the CCP and the clearing member is detected in intraday mode, the anomaly is more likely to be explained by a data quality issue.

Example 2: concentration. The second example uses the ADQ framework to disentangle the anomaly downstream. Let us suppose we observe a discrepancy between the exposures (proxied by notional amounts) reported by two EU counterparties (A and B). Running the ADQ in intraday mode suggests that the issue is due to missing contracts from counterparty A. In this case, any further anomaly including data reported by counterparty A is more likely to be due to data quality issues, rather than being a financial stability signal.

6 Conclusions

Despite policymakers' efforts in collecting granular level data after the global financial crisis, persistent and pervasive data quality issues increase opacity, thereby hampering the ability to analyse data and produce effective policy responses. This paper describes a novel framework to identify factors underlying the developments in large, granular datasets of financial information and proposes several applications based on data on derivatives collected under the EMIR Regulation. We show how tools build on this framework have been successfully deployed on the ECB IT infrastructure, and are regularly used to decompose the changes in certain measures of derivative markets into data quality issues and genuine developments that may have potential impact on the financial stability.

One of the essential features of the tool is its customisability, allowing the relevant staff to apply the solution to various datasets, measures, and dimensions, as well as employ any initial filtering deemed necessary. Thus, while the original application was data collected under EMIR, we plan to apply the tool to other granular datasets available at the ECB and ESRB.

Given the structured output of the tool, this daily process can be further integrated into other daily monitoring operations on the granular datasets, significantly reducing efforts needed to ensure that the information ingested is correct, allowing also for considerable reduction of the time needed for identification and reporting of data quality issues.

References

Abad J., Aldasoro I., Aymanns C., D'Errico M., Fache Rousová L., Hoffmann P., Langfield S., Neychev M., Roukny T. (2016). Shedding light on dark markets: First insights from the new EU-wide OTC derivatives dataset. ESRB Occasional Paper Series No 11, 2016.

Apicella E., Ciullo A., Übelhör C., Marques P., D'Errico M. (2023). Monitoring *at scale*. Forthcoming.

Brunnermeier, M.K., Gorton, G. and Krishnamurthy, A., 2012. Risk topography. NBER Macroeconomics Annual, 26(1), pp.149-176.

Carraro T., Fache Rousová L., Furtuna O., Ghio M., Kallage K., O'Donnell C., Vacirca F, Zema S. M. (2021). Lessons learned from initial margin calls during the March 2020 market turmoil. *ECB Financial Stability Review*, November 2021

Du D. (2018). Essential Techniques to Help You Process, and Get Unique Insights from, Big Data, 2nd Edition. Packt Publishing, Limited.

Duffie D. (2011). A 10-by-10-by-10 Approach. https://www.darrellduffie.com/uploads/policy/Duffie10By10By10July2011.pdf

ECB Banking Supervision (2018). Report on the Thematic Review on effective risk data aggregation and risk reporting. https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.BCBS_239_report_201805.pdf

Erl T., Khattak W, Buhler P. (2016). Big Data Fundamentals: Concepts, Drivers & Techniques. Pearson

ESMA (2021). EMIR and SFTR data quality report 2020, https://www.esma.europa.eu/sites/default/files/library/esma80-193-1713_emir_and_sftr_data_quality_report.pdf

ESMA (2020a). Final Report. Technical standards on reporting, data quality, data access and registration of Trade Repositories under EMIR REFIT

ESMA (2020b). Report to the European Commission on post trade risk reduction services with regards to the clearing obligation under EMIR Article 85(3a). https://www.esma.europa.eu/sites/default/files/library/esma70-156-3351_report_on_ptrr_services_with_regards_to_the_clearing_obligation_0.pdf

ESMA (2017). Final Report – Guidelines on transfer of data between Trade Repositories. https://www.esma.europa.eu/sites/default/files/library/esma70-151-552_guidelines_on_transfer_of_data_between_trade_repositories.pdf

European Systemic Risk Board (2022). ESRB's view regarding data quality issues and risks for financial stability. https://www.esrb.europa.eu/pub/pdf/other/esrb.letter220713_on_data_quality_issues~18eccb6993.en.pdf

European Systemic Risk Board (2020a). Liquidity risks arising from margin calls. Available at: https://www.esrb.europa.eu/pub/pdf/reports/esrb.report200608_on_Liquidity_risks_arising_from_margin_calls_3~08542993cf.en.pdf

European Systemic Risk Board (2020b). Secretariat staff's response to ESMA's consultation paper on technical standards on reporting, data quality, data access and registration of trade repositories under EMIR Refit. https://www.esrb.europa.eu/pub/pdf/other/esrb.letter200812_response_to_ESMAs_consultation_paper~bef2263d90.en.pdf?0b34782ea7527a3e8a322cb3b124c097

FSB, IMF (2009). The Financial Crisis and Information Gaps, Report to the G-20 Finance Ministers and Central Bank Governors. https://www.fsb.org/wp-content/uploads/r_091029.pdf

Lai R., Potaczek B. (2019). Hands-On Big Data Analytics with Pyspark : Analyze Large Datasets and Discover Techniques for Testing, Immunizing, and Parallelizing Spark Jobs. Packt Publishing, Limited.

Loshin D. (2010). The Practitioner's Guide to Data Quality Improvement. Morgan Kaufman

Mahanti R. (2019). Data Quality: Dimensions, Measurement, Strategy, Management, and Governance. ASQ Quality Press

Nasiriany S. Thomas G., Wang W., Yang A. (2019). A Comprehensive Guide to Machine Learning. https://www.eecs189.org/static/resources/comprehensive-guide.pdf

Sambasivan N., Kapania S., Highfill H., Akrong D., Paritosh P., Aroyo L. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. https://research.google/pubs/pub49953/

EUROPEAN CENTRAL BANK
EUROSYSTEM

Anomaly intersection: disentangling data quality and financial stability developments in a scalable way

Gemma Agostoni (ECB)
Louis de Charsonville (McKinsey & Company)
Marco D'Errico (ECB, ESRB Secretariat)
Cristina Leonte (BIS)
Grzegorz Skrzypczynski (ECB)

The views expressed here are of the authors and do not necessarily represent the views of the associated institutions

IFC-Bank of Italy workshop
Data Science in Central Banking: Applications and tools
15 February 2022

Background

- Following the financial crisis of 2008/2009, policymakers now have access to several large scale & granular–level datasets, implying the need to scale up monitoring and analytical work
- However, persistent and pervasive data quality issues (largely due to reporting agents and trade repositories) hamper this process, reducing transparency
- Policymakers are now facing a double challenge: how to disentangle developments that are relevant from a financial stability perspective from those resulting by bad data quality?

Quality in datasets of granular financial information

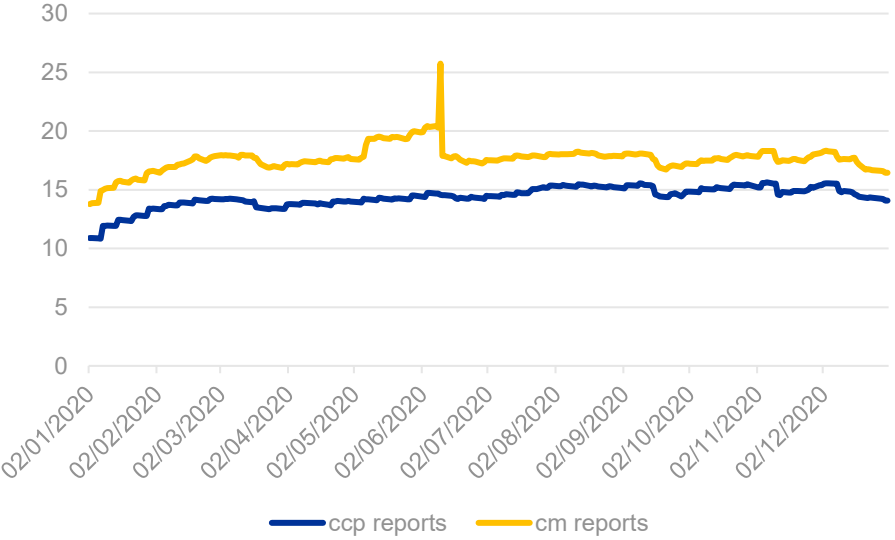
The financial crisis of 2008/2009 led to implementation of multiple high-frequency collections of granular financial information.

AIFMD
STS
Banks' large exposures
AnaCredit
SFTR
EMIR

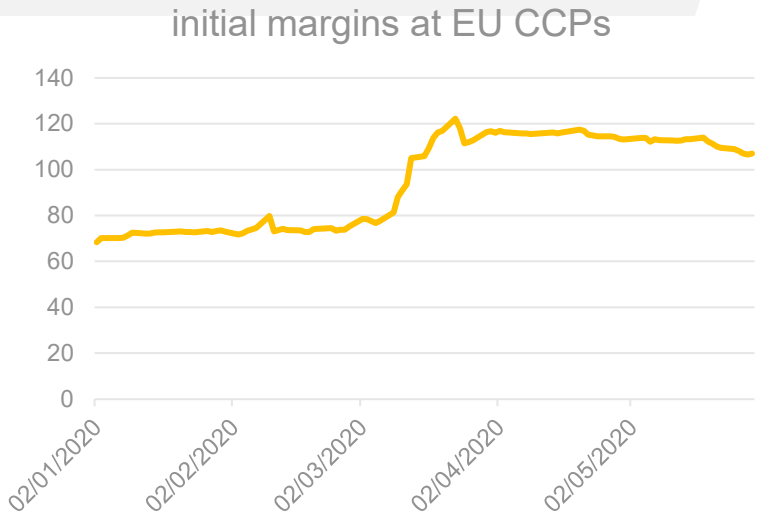
Those dataset pose a unique challenge to the regulators due to their enormous size and insufficient data quality

IT and TR issues	Misreporting
5 million duplicated trades sent daily over a month	Inconsistent information reported by CCPs and clearing members
Negative values incorrectly changed to absolute values for 1.5 years	Incorrect signs of contract values
Missing collateral reports (IM + VM) for large CCPs	Not following the reporting guidelines (e.g. collateral portfolio code, fx swaps)
Information reported by counterparties not passed onto the reports for authorities (e.g. "Asset class", "Contract type", collateral variables)	Implausible numerical values (reaching EUR trillions) – also by CCPs and other large entities
Disappearing negative rates	CCPs reporting no outstanding positions at end-day

Disentangling data quality and financial stability developments

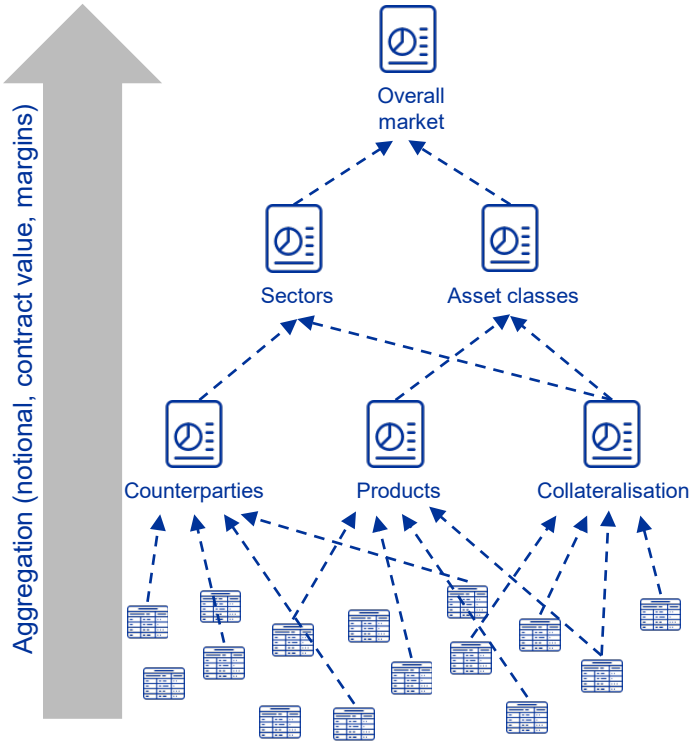


Notional amounts between EU CCPs vis-à-vis 1200+ EU clearing members differ substantially -> data quality issue



Is the dramatic increase in initial margins at EU CCPs during the March 2020 turmoil a development or a data quality issue?

Bridge between micro and macro



Granular data like EMIR blurs the line between macro and micro – **the final users can seamlessly zoom-in and zoom-out across aggregation levels** from analysing individual trades to assessing the overall market.

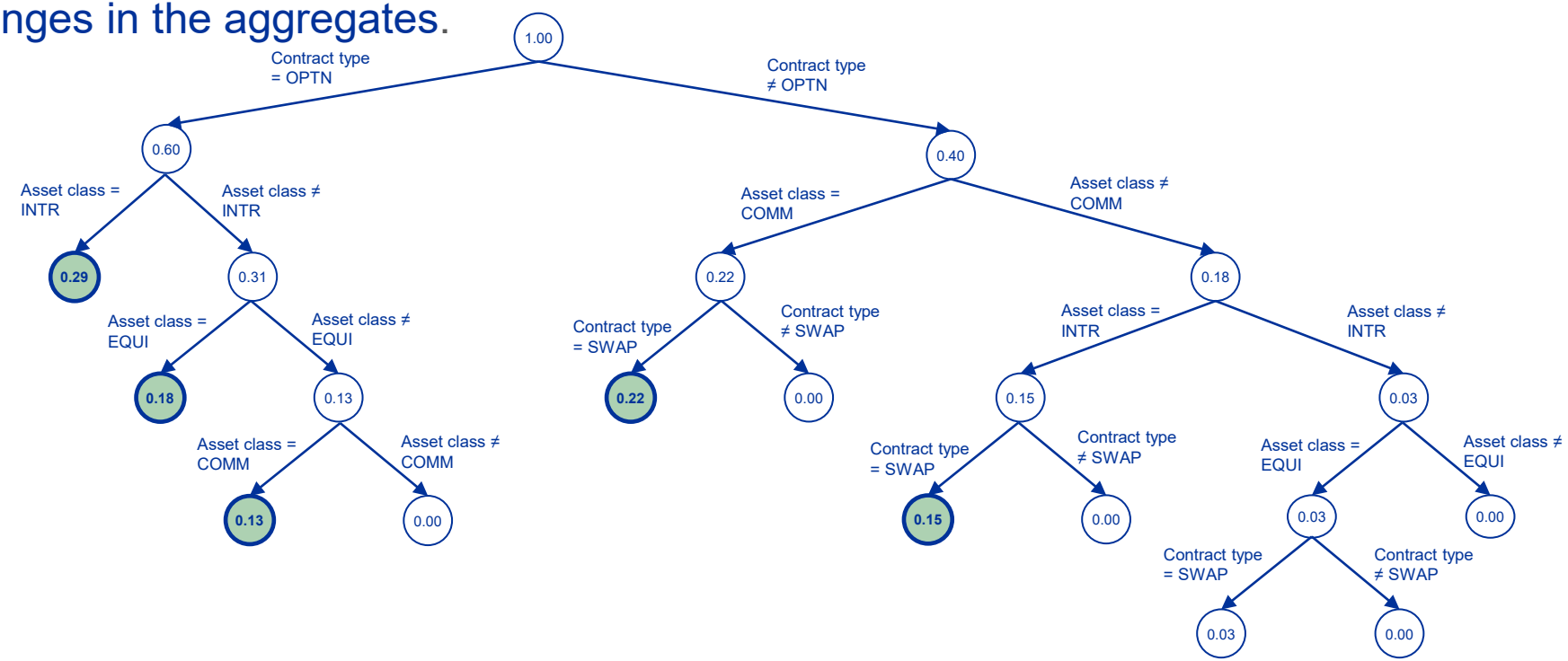
But trying to apply **traditional manual or semi-automatic tools** to data quality management is like looking for a **needle in a haystack**.

There's **no time** to laboriously look for the answers, when something unusual happens.

We want **the answers** to wait for us **in our mailboxes every morning** – before we even ask the question!

Dimension analysis

We use decision trees to determine dimensions that best characterize the changes in the aggregates.



Entity-level analysis

- **Aggregate the measure** reported by the relevant entities
- **Compare the values** reported in two reference periods analysed
- **Select the entities** with biggest impact on the change in the measure

t

Entity	d^1	...	d^U	Notional
ID1	A		X	100
ID2	A		X	40
ID2	B		Y	20
ID3	B		Z	20
ID3	A		Z	50

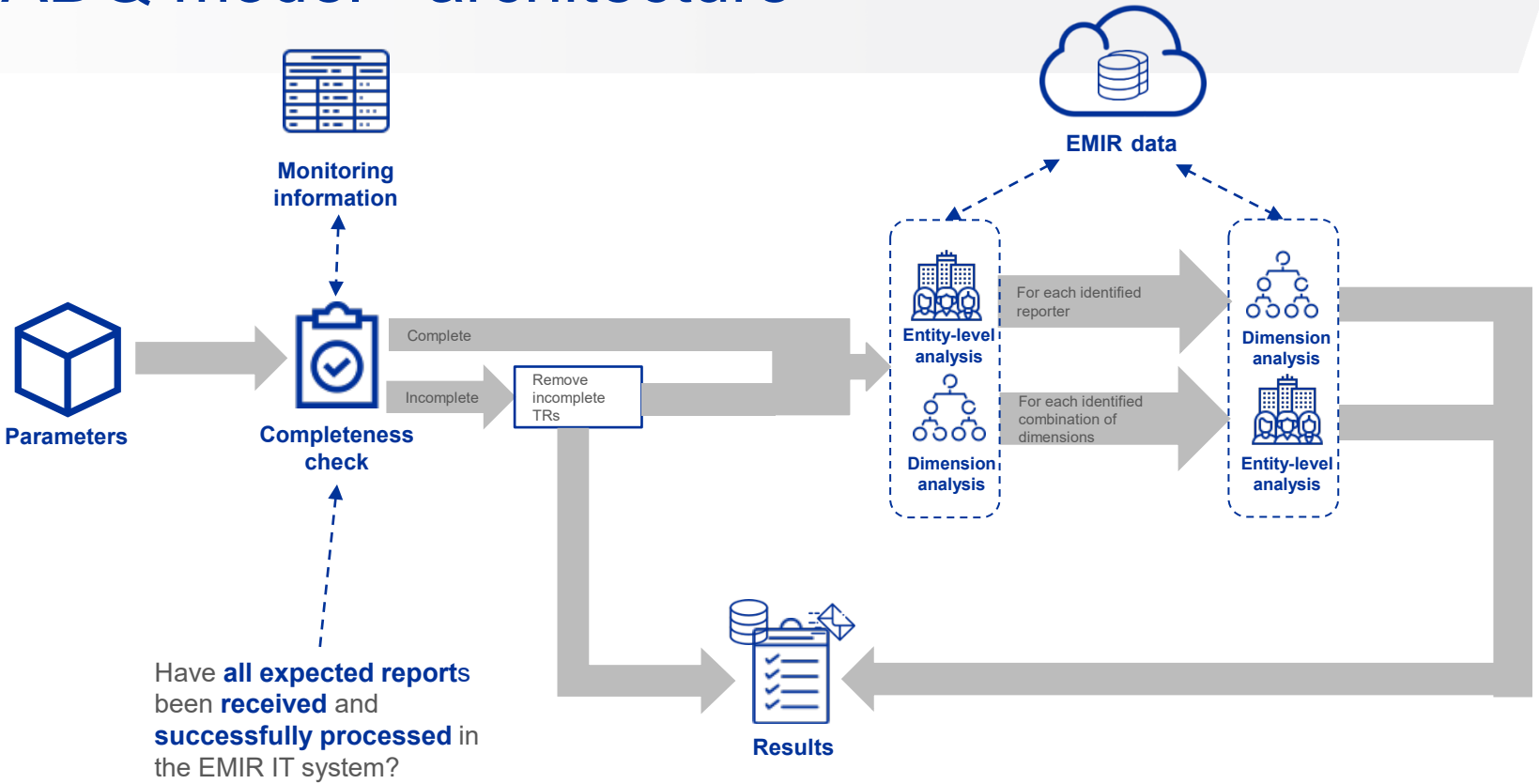
t-1

Entity	d^1	...	d^U	Notional
ID2	A		X	50
ID2	B		Y	70
ID3	B		Z	10
ID3	A		Z	110
ID3	A		X	40

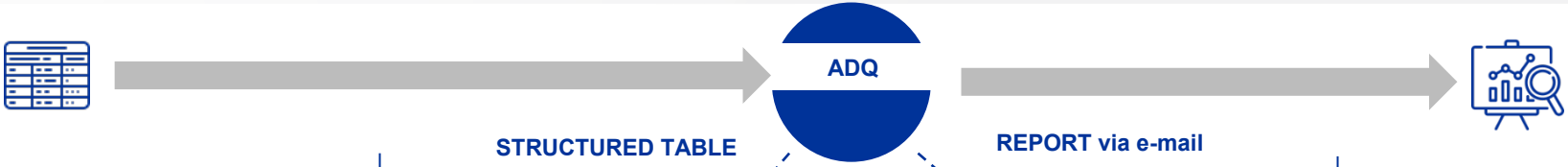
Entity	$Notional_t$	$Notional_{t-1}$	Δ	$\lvert\Delta\rvert$
ID1	100	0	100	100
ID3	70	160	-90	90
ID2	60	120	-60	60

The model allows the reporters to be treated as dimensions, however, given the number of unique reporters this would significantly reduce the performance

ADQ model - architecture



ADQ workflow & features



Parameters

- **Measures**
- **Dimensions**
- **Concentration threshold**
- **Type of analysis (time-series vs. intra-day)**
- **Input datasets**

- **Multiple jobs** running every day reflecting different types of analysis
- Each job is defined by **a separate parameters' set**
  - **Time-series** analysis of **notionals**, **contract values**, and **initial margins**
  - **Intra-day** analysis of **notional** reported by **CCPs** and **CMs**
- **Extension** to other datasets on-going (SFTR)

Follow-up

- **Further investigation** by experts
- **Transmission** of identified issues to **ESMA**
- Informing **internal users**

Conclusions & way forward

ADQ – main features

- Timeliness
- Flexibility
- Analytical support
- Dataset agnosticity

Way forward

- Supporting the analysis with information from **activity reports** (flow)
- Building **complex, configurable pipelines**
- **Semi-automated transmission** of issues to **ESMA**

THANK YOU FOR YOUR ATTENTION