Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Data science in central banking"

18-20 February 2025

Data anonymization principles at Banco de Portugal[1]

Ana F Carvalho, Francisco Fonseca, Mário Lourenço
and Ricardo Marques,
Banco de Portugal

[1] This contribution was prepared for the workshop. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Data anonymization principles at Banco de Portugal

Ana F. Carvalho[1], Francisco Fonseca[2], Mário Lourenço[3], Ricardo Marques[4]

Abstract

Central Banks collect increasingly larger sets of granular data which frequently contain personal data. The analytical potential of these datasets is greatly enhanced whenever they have common identifiers, meaning they can be combined in order to maximize the value of the available information. However, given the sensitive nature of some of these datasets, particularly when referring to natural persons, the protection of personal data must always be a priority. This paper focuses on the essential principles that must be observed when defining a robust anonymization model aimed at preserving the confidentiality of individual data, while still allowing the use of different combinations of the available datasets for analytical purposes. We focus on pseudonymization mechanisms and the definition of different access profiles as cornerstones of a centralized anonymization policy for Banco de Portugal's databases.

[1] Banco de Portugal, Statistics Department – afcarvalho@bportugal.pt.

[2] Banco de Portugal, Statistics Department – ffonseca@bportugal.pt.

[3] Banco de Portugal, Statistics Department – mfllourenco@bportugal.pt.

[4] Banco de Portugal, Statistics Department – rjlmarques@bportugal.pt.

The opinions expressed are those of the authors and do not necessarily coincide with those of Banco de Portugal or the Eurosystem.

Contents

1. Introduction

Data protection is of paramount importance to every organization. Dealing with the risks associated with the possible wrongful disclosure of data each organization is entrusted with is not, however, straightforward.

Central banks, due to the nature of their activity, have access to increasingly vast sets of data reported directly by economic agents or obtained indirectly through other sources. Efforts to reduce the reporting burden of economic agents have contributed to the wider use of granular data which, single-handedly, allow respondents to fulfil different reporting requirements. The growing availability of granular datasets, associated with the computational capabilities which enable its analysis using self-service tools, have led to the widespread access and use of microdata.

These trends impose the need for the effective implementation of data protection mechanisms, particularly given the sensitive nature of datasets that refer to microdata concerning natural persons. In such cases, the protection of personal data must be upheld at all costs. Nonetheless, given that the analytical potential of these datasets is greatly enhanced whenever they have common identifiers, it is essential to define a strategy that effectively weighs the advantages and disadvantages between virtually indiscriminate access to all available data (enhanced by self-service tools) and its effective protection. This is especially true considering organizations need to address the relevant procedures to comply with recent regulation developments on this topic (namely the General Data Protection Regulation or "EU GDPR"[5]). The discussion regarding the adoption of IT infrastructures based on cloud solutions also plays a role regarding these issues. Ideally, we should seek to minimize the impact that the implementation of constraints on data accessibility has on the analytic potential of that data. It is up to every organization to define, in the context of their respective data management strategies, how they can address these concerns and to what extent it is possible to reconcile these forces.

This paper focuses on the essential principles that must be observed when defining a robust anonymization model aimed at preserving the confidentiality of individual data, while still allowing the use of different combinations of the available datasets for analytical purposes. We focus on pseudonymization mechanisms and the definition of different access profiles as cornerstones of a centralized anonymization policy for Banco de Portugal's databases. Section 2 provides an overview of the recent developments within Banco de Portugal's data infrastructure and how they are linked with the need to enhance data protection procedures. Section 3 presents the basic principles of an anonymization strategy set out to be the backbone of this data infrastructure, while Section 4 describes a set of recommendations outlined for the definition and governance of an anonymization model to be transversely applied at Banco de Portugal. Section 5 presents our closing remarks and sets out the way forward towards the implementation of this model.

2. Recent developments regarding Banco de Portugal's data infrastructure

The availability of new and more granular data on a variety of subjects, directly reported by different sets of economic agents, obtained through administrative data sources, or via the implementation of a multitude of data collection mechanisms, has led to the widespread conception that relevant data is virtually everywhere we look. This is particularly true in the case

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 of April of 2016.

of Central Banks. With more data on various aspects of the business landscape being used for several purposes (supervision, statistical production, economic analysis, among others), the need to invest in robust data infrastructures and implement appropriate information management principles (and practices) has become evident.

Data sharing has become particularly relevant for any organization seeking to use its data more efficiently. However, sharing data comes at a cost, as protecting the data available at any organization while making it readily available and useful across different teams, and for a variety of purposes, requires balancing data users' needs and the legal and ethical obligation of protecting data confidentiality.

This is particularly true when we consider granular data on natural persons. Accessing such data (behavioral data, financial data, etc.) has great potential for several purposes. Its exploration has been enhanced by the use of self-service tools with increasing computational capabilities which are widely available at any organization. Recent developments on artificial intelligence and machine learning algorithms have also led to new ways of using already available data.

But any dataset (particularly data on natural persons) needs to be handled carefully. The wrongful disclosure of data at the disposal of any organization, directly or indirectly obtained, not only poses legal risks (EU GDPR introduced restrictions on the use and storage of data relating to natural persons, admitting, however, its use for scientific and statistical research purposes) but also reputational risks, as it can undermine confidence on the capability organizations have to deal with the data they are entrusted with (which is of particular relevance in the case of Central Banks, for instance). Data security risks need to be accounted for at all costs, specifically bearing in mind recent developments on data infrastructures and the transition to cloud-based data management solutions.

At Banco de Portugal, the recognition of the need to address these challenges led to the implementation of an Integrated Data Management (IDM) Program, which was defined as part of Banco de Portugal's Strategic Plan for the 2017-2020 period, focusing specifically on providing an information landscape that enables practices aimed at maximizing the potential of available data, rationalizing data collection and data production processes, while also determining the need for data to be increasingly shared to fulfil different business needs (MORENO, 2021). The most recent Strategic Plan (for the 2021-2025 period) established the need to consolidate the IDM Program and to evolve the relevant data governance models and data architecture, aimed at turning Banco de Portugal into an organization focused on data.

Centralized data repositories and harmonized data processing procedures, including data quality management routines, are, nowadays, cornerstones of a data driven culture where information is available across the institution. A new data architecture based on a set of these fundamental principles (standardized processes, common data dictionaries, shared data repositories with access control and auditability, but where virtually all data is at the disposal – if needed – of every member of the institution, etc.) was envisaged as part of a streamlined conceptual framework which Banco de Portugal's Statistics Department has been progressively implementing (GONÇALVES et. al., 2023), also aiming at upscaling this experience for other data domains. Banco de Portugal has recently begun to design the BdP DataHub, a project that aims to concentrate and streamline potentially all data and data related processes at the institution (regarding statistical data, supervisory data, payments data, etc.) in an effort to increase the efficiency of our information management procedures, reducing data redundancies and making data available for whatever purposes analysts see fit.

These strategic decisions, nonetheless, need to be balanced with the increasing availability of particularly sensitive datasets (on loans granted by financial institutions, on banking deposits, on household income, etc.) which need to be protected. For this, a clear access policy is essential, ensuring that sensitive data is available only on a need-to-know basis (for quality control

procedures or for supervisory tasks, for instance). For all other purposes, further steps need to be taken to protect data confidentiality.

With these aspects in mind, it became clear that Banco de Portugal's data management strategy also had to consider the viability of implementing a set of anonymization techniques and procedures that would counter-weigh the growing availability (and need) of data, building upon previous experiences in order to minimize the loss of data integrity while maintaining data security. The following Sections describe the set of principles which are being discussed as the bases of Banco de Portugal's anonymization strategy.

3. Anonymization principles

The definition of an anonymization strategy within an institution such as a Central Bank relies, primarily, on the need to define clearly a set of concepts that should be commonly understood by virtually all data experts: what is or is not considered to be personal data, what is the focus of the common anonymization strategy, what are the perceived differences between anonymization strategies and the implementation of pseudonymization techniques. These are all fundamental concepts that need to be tackled.

With these aspects in mind, a set of key issues then need to be identified and tackled when prototyping a solution – should the solution be implemented across the institution or within a given data domain, should a public pseudonymization algorithm be used or should an internal solution be tested, what are the constraints related with the implementation by a given business unit and what should the expected data flow be when dealing with datasets that should be anonymized, among others. We will tackle these issues mostly in section 4.

3.1. Main concepts and definitions

The type of information that usually is considered relevant within the context of data protection is commonly called personally identifiable information (PII) or, more informally, personal data. Personal data is any piece of information that allows data users to determine, directly or indirectly, the identity of a certain agent (usually natural person, although the same reasoning may be considered for any legal person). The assessment of what is or is not personal data must be as objective and devoid of personal judgments as possible.

PII can be categorized into two types: i) identifiers, i.e., data that, by itself, allows us to uniquely identify a specific individual (or economic agent) – for example, an ID number – or; ii) quasi-identifiers, i.e., data that, when cross-referenced together with other data allows the identity of an agent to be indirectly inferred – for example, the first name, address and age of an individual are characteristics that may not directly identify that individual, but that may, when combined, allow data users to infer the individual's identity with a high degree of confidence.

The second type of PII is particularly relevant when we consider rare combinations of some personal data characteristics. Assessing the re-identification risk associated with a given dataset is not a linear task, implying the need to analyze the context of a given analysis or combination of datasets (i.e., the database, or even a set of databases). This implies that the anonymization of a database is equally non-linear in many cases, a problem that is enlarged when we try to define an anonymization solution that can deal with the need to cross-tabulate different databases which, individually, may present low risk of identification, but, when combined, represent a greater risk.

Data anonymization mechanisms are processes that make it possible to eliminate or, at least, significantly reduce the identifying nature of PII. In general, this can be achieved through three broad families of anonymization algorithms:

i) Elimination-based algorithms – These algorithms remove or replace with a meaningless value the dimensions of the data that present a high risk of identification.

ii) Replacement-based algorithms – These methods replace values in high-risk dimensions with dummy values (pseudonyms) that preserve the original format, uniqueness, and data type.

iii) Aggregation-based algorithms – These techniques aggregate the values of high-risk dimensions to reduce the risk of identification.

The three types of techniques can also be combined to increase robustness. However, there is generally a strong positive relationship between the reduction of the risk of identification and the implicit loss of information.

A key distinction that is also relevant, particularly in our case, is the difference between anonymization and pseudonymization. Under European regulation, data is anonymized if and only if it is altered in such a way that it is no longer possible, under any circumstance (even with additional data), to restore the original data (i.e., re-identify any individual). In essence, this means that the underlying anonymization algorithm must be irreversible, while also ensuring that inference (with a given degree of confidence) is not possible. On the other hand, pseudonymization allows for reversibility under specific circumstances.

In the remainder of the text, we will use the term "anonymization" to refer to any given method that aims to reduce the risk of re-identification, regardless of its reversibility.

Finally, any anonymization model consists of three key pieces: (1) identified information; (2) anonymization mechanism; and (3) anonymized information. Applying an access policy that considers each of these components is a key factor to ensure its viability. This is particularly relevant in the case of reversible pseudonymization algorithms, since if a given user has access to two out of three information pieces, then the remaining piece is easily derived.

A robust data anonymization mechanism, potentially combining different techniques, coupled with a strong data access policy that limits the knowledge different teams have about the datasets available within the organization, are key aspects to bear in mind when defining and implementing an effective anonymization/pseudonymization framework.

4. Defining the path towards implementation

The data anonymization framework discussed in this paper focuses on four essential aspects:

- The identification of relevant databases and high-risk variables;
- The definition of an access policy which takes into account different sets of stakeholders;
- The assessment, implementation and governance of a standardized anonymization mechanism ; and,
- The need to define and implement policies governing every datasets' entire lifecycle within the organization, encompassing acquisition, transformation and sharing.

The following sections detail how we envisage the implementation of a data protection framework at Banco de Portugal considering these aspects.

4.1. Identifying relevant set of databases and critical variables

Any anonymization policy should necessarily include a comprehensive list of databases and variables that, individually or when combined, may result in significant identification risk (essentially, a data catalogue). Along with this assessment, one should also evaluate the relevance of these high-risk variables for analytical purposes within the organization, which will necessarily impact the anonymization mechanism that should be implemented. As an example, an individual's name is, in general, of very little interest for most users in the context of the analytical exploration of any database, meaning that such a variable can likely be entirely removed from pseudonymized datasets. On the other hand, an individual's age, or nationality, which also carry significant re-identification risk, will likely be relevant for certain purposes, meaning that they will need to be handled more carefully (for example through aggregation).

It is thus difficult, in general, to find a balance between reducing the risk of re-identification and keeping the data useful for analytical purposes. For this reason, it might be justifiable to create an internal team whose sole purpose is to analyze data requests from users, anonymize the underlying data (if necessary), and grant access to the anonymized dataset to the subset of users that requested it, as this is the only way to ensure that the trade-off between privacy and analytical potential is optimal for each specific case. Summarily, an information request would encompass the following steps:

i) User(s) request access to database(s). The request must specify the databases to be accessed, and the variables of interest.

ii) The designated team assesses the request to determine if it includes personal data, and identifies the high-risk variables involved. It then selects the appropriate anonymization procedures needed to achieve an acceptable level of re-identification risk. If the request involves data from multiple databases, the assessment must consider if they have common keys which allow them to be joined. In such cases, re-identification risk must be evaluated for both the individual databases and the combined dataset. Methods like k-anonymity and l-diversity can aid in this assessment.

iii) The assessment team determines if an acceptable risk level can be achieved and performs the anonymization procedures.

iv) The assessment team shares the anonymized data with the user(s) that requested it.

To further aid in this assessment, high-risk variables can be categorized by risk level – for example, an individual's date of birth poses a higher re-identification risk (in general) than their age. This approach would provide the assessment team with a good starting point for their analysis. In fact, the Banco de Portugal Microdata Research Laboratory (BPLIM) already provides a similar service, targeted mostly at providing external researchers with access to anonymized micro datasets, although internal researchers can also make this kind of request. Hence, this internal team would essentially provide this service for all internal users, and for a wider range of datasets.

However, this approach would have significant resource allocation costs, especially since the number and complexity of information requests are likely to increase over time. Furthermore, it is highly likely that this approach would result in significant redundancies between different data requests, meaning the team would need to analyze very similar requests repeatedly, often granting access to very similar datasets. For these reasons, we consider that the implementation of a solution that is less robust but more agile and flexible may be desirable.

To this end, the algorithm to be implemented would need to focus on the anonymization of direct identifiers (such as ID numbers and full names), as these variables must always be

transformed, regardless of context, due to their inherently high re-identification risk. To complement this "baseline anonymization", applied to every relevant database, each of them would also be analyzed in isolation regarding the re-identification risk presented by the remaining variables. Summarily, the process would encompass the following steps:

i) A designated team assesses each database to determine if it includes personal data and identifies the high-risk variables. It then applies a harmonized anonymization procedure targeted at any direct identifiers. The team then determines if any further procedures are justified based on the database's re-identification risk (aided by the methods mentioned above).

ii) The team determines if an acceptable risk level can be achieved and performs the additional anonymization procedures.

iii) The assessment team shares the anonymized data with all eligible users (the guidelines for eligibility will be addressed in section 4.2).

This solution, although not as effective as the first at reducing re-identification risk, avoids the proliferation of redundant data access requests, and would be much less costly in terms of resources, while still significantly enhancing data privacy at an internal level.

4.2. Implementing a robust data access control policy focused on protecting data

An anonymization policy must include the definition of access profiles that reduce the risk of the anonymization process being undermined.

As mentioned previously, an anonymization model encompasses three key pieces: (1) identified information; (2) anonymization mechanism; and (3) anonymized information. In the case of reversible anonymization mechanisms, knowing two of these pieces allows a user to easily derive the third one. An effective access policy must, therefore, forbid the vast majority of users from accessing more than a single piece.

To this end, we propose the existence of three distinct access profiles:

i) Type A access is granted to users that require access to identified datasets to perform their work – functions such as quality control, systems administration, and supervisory tasks might warrant this type of access. Furthermore, each user will only have access to the minimal set of information they require.

ii) Type B access is granted to users responsible for conducting anonymization procedures. These users are responsible for the implementation and maintenance of the anonymization pipeline, and thus must have access to all identified data.

iii) Type C access is granted, in general, to all users that do not have Type A or B access. Type C users can access all anonymized datasets and will use them for the purposes of statistical production or analytical exploration.

It is also relevant to keep in mind that, to ensure the effectiveness of the model, any user with Type A access cannot have Type C access, and vice-versa. Furthermore, Type A and Type C users cannot have access or knowledge of the anonymization mechanism. In the case of Banco de Portugal, it is also relevant to say that this type of access policy would require significant changes regarding the daily routine of teams, since many of them perform tasks related both to quality control (where access to identified data is necessary) and to statistical production (where this type of access is not necessary).

4.3. Defining an anonymization mechanism

In this section, we will be focusing mostly on a proposal regarding an anonymization mechanism targeted at pseudonymizing direct identifiers. It is also relevant to keep in mind, as a starting point, that the anonymization procedures that were considered must transform each identifier into a unique pseudonym. This ensures that if two datasets share a common identifier and can be combined, then the anonymized datasets will also be joinable. Furthermore, pseudonymization is considered an appropriate safeguard for personal data within the context of the EU GDPR.

The mechanisms being proposed replace direct identifiers with pseudonyms, preserving uniqueness, and in some cases also data type and length (i.e., a numeric identifier would have a numeric pseudonym with the same number of digits). A relevant characteristic of pseudonymization algorithms is that they are reversible, although their reversibility should only be envisaged when strictly necessary, while also resulting in minimal (in most cases entirely negligible) loss of information.

There is a wide range of possible pseudonymization algorithms that could be adopted, of which we can highlight:

i) Hash functions, that map identifiers to a hash value (which in most cases has fixed length) and are irreversible. In order to preserve reversibility, a mapping table must be kept and updated in parallel.

ii) Transposition and substitution cyphers, which are methods that scramble characters or replace them with different ones, respectively.

iii) Symmetric and asymmetric-key based algorithms.

Furthermore, these algorithms can be combined or iteratively applied in order to increase complexity and security. There are also key-based hash functions and cyphers that, in general, will be more secure than the baseline methods, which are not secure in cases where we know the length of the input – as an example, a US social security number is a random 9 digit number, and if it were pseudonymized using a method like standard SHA-256 hashing, it would be trivial for any person to simply generate hashes for all possible 9 digit numbers, establishing a 1-to-1 mapping between any number and the corresponding hash. Thus, we believe that a unique combination of a set of different key-based algorithms would be the best way to achieve an acceptable level of security. This unique algorithm would then be applied as a baseline to all datasets containing direct identifiers.

Naturally, this mechanism will then need to be complemented with an assessment of the risk associated with the remaining variables within a given dataset, in order to determine if additional measures must be taken in order to achieve an acceptable level of re-identification risk.

4.4. Establishing a common data flow

The effectiveness and robustness of any anonymization model depends on the knowledge and proper application of the data flow defined to protect data confidentiality at any given institution. The exchange of data for ad-hoc analyses, for instance, can jeopardize the effectiveness of such a model. As so, any set of data reported to or obtained by Banco de Portugal should be evaluated and treated according to the defined data flow, which should include the application of the established anonymization model when relevant. This implies that the operationalization of the model must ensure that it is flexible and efficient enough for users to effectively go through the process (minimizing the possibility of jeopardizing the data flow by sharing information outside its scope).

Although data is reported and received through a variety of different channels and using different file formats, we will focus on two general cases – data reported to Banco de Portugal directly by reporting agents, and data received indirectly from other entities.

It should be noted that, regardless of which case we are talking about, it must be ensured that the defined data flows must be versatile and flexible enough to safeguard that the data is made available to the final user in a timely manner.

4.4.1. Data reported directly

Directly reported data is, in most cases, fully identified – for example, information received regarding loans to natural persons includes the individual's tax identification number, as well as a multitude of personal quasi-identifiers (such as name, gender, and date of birth). Furthermore, this information will generally have to be subjected to a multitude of data quality control checks, some of which require unrestricted access to the original data.

Thus, upon being received, the identified data will be made available to the smallest possible subset of users (users with Type A access, as defined in section 4.2), who will perform the necessary data quality checks. For all other purposes, the data will be subjected to the defined pseudonymization mechanism (conducted by Type B users), where, as a minimum, all direct identifiers will be pseudonymized. If the dataset still poses significant reidentification risk, further measures may be appropriate. After an acceptable risk level is achieved, the pseudonymized dataset can then be made available to Type C users.

4.4.2. Data received from third parties

When the data received from third parties is fully identified, the data flow to be followed is mostly the same as the one for directly data reported, with the notable exception of the need to perform data quality checks, as we assume that, in most cases, this data has already undergone strict quality checks. Thus, it may be unnecessary for any user to have access to the original data, and it can immediately undergo the necessary pseudonymization procedures.

However, there are cases where the data received from third parties is already pseudonymized. In such cases, it is necessary to keep in mind that if the intent is to join this data with other datasets available at Banco de Portugal, then it is mandatory that the third party also shares the underlying pseudonymization algorithm. In such a case, it is then possible for Type B users to pseudonymize the necessary internal data with the same algorithm, thus ensuring that the pseudonymized datasets can then be joined by Type C users.

5. Closing remarks

Defining a robust anonymization model aimed at preserving data confidentiality is not a simple task on its own, let alone when we try to protect sensitive data while still allowing for it to be used for different purposes. Having the best of both worlds is particularly difficult.

With our work we tried to establish a set of principles which will enable the implementation of an anonymization model aimed at implementing effective data protection procedures without jeopardizing recent developments regarding data sharing within our organization. Our recommendations target the need to identify relevant databases and high-risk variables, define an access policy which takes into account different sets of stakeholders, implement a standardized

anonymization mechanism, while having to define and implement policies governing every dataset's entire lifecycle.

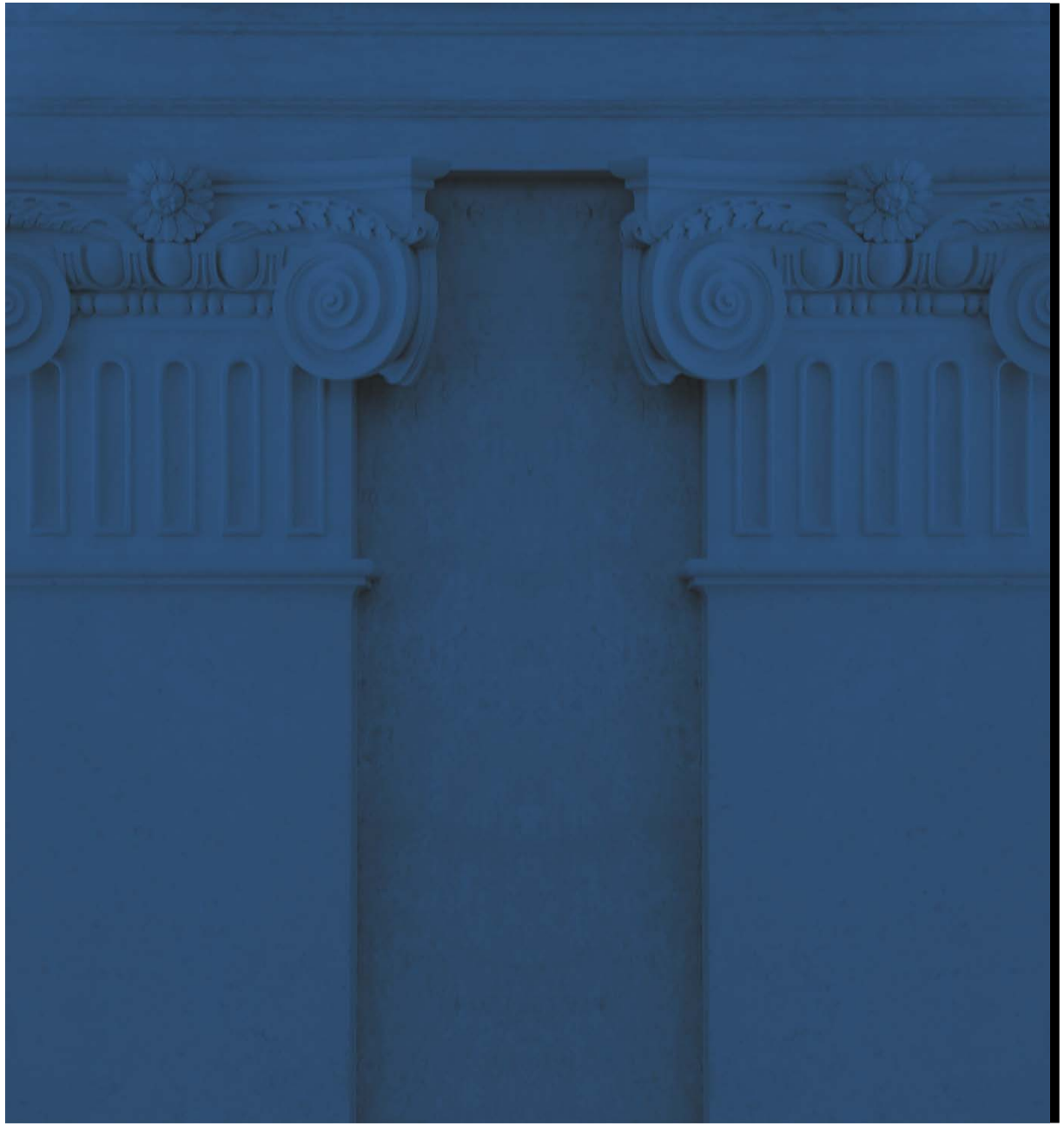
These recommendations need, nonetheless, further validation. We are now at a point where we are discussing if they are in fact viable, particularly in what concerns the segmentation of roles and responsibilities. If these recommendations introduce too much "noise" regarding the way people already access and explore data, they may need to be reassessed and other alternatives may need to be explored. Working on a number of use cases where we apply pseudonymization techniques to different databases, making them available to specific sets of users for analytical exploration, will bring valuable feedback on whether this model is in fact feasible, and what are the issues we need to address for us to have an harmonized way to protect our datasets.

6. References

GONÇALVES, Ana R., et. al. (2023), New strategy of data sharing and data access in statistics: the view from Banco de Portugal, presentation at the 3rd IFC Workshop on Data Science in Central Banking (Rome, October 2023).

MORENO, M. Carmo (2021), Data Governance: an orchestra of people, processes, and technology, in IFC Bulletin No 54 (2021), Issues in Data Governance.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 of April of 2016.



DATA ANONYMIZATION PRINCIPLES AT BANCO DE PORTUGAL

ANA F. CARVALHO
FRANCISCO FONSECA
MÁRIO LOURENÇO
RICARDO MARQUES

4TH IFC-BDI WORKSHOP ON
"DATA SCIENCE IN CENTRAL BANKING“

ROME, FEBRUARY 2025

BANCO DE PORTUGAL
EUROSISTEMA

AGENDA

01 | MOTIVATION

02 | ANONYMIZATION PRINCIPLES

03 | DEFINING THE PATH TOWARDS IMPLEMENTATION

04 | CLOSING REMARKS

MOTIVATION

01

?

- How do we balance the need to **protect our data** while keeping it available to be easily **used for different purposes**?

MOTIVATION
GRANULAR DATA AND DATA PRIVACY

Granular data containing information on natural persons, which is increasingly available...

... Has great potential

- Self-service exploration for analytical purposes
- Elimination of redundancies which may lead to a reduction of the reporting burden of institutions
- Making full use of increasing computational capabilities and **ML/AI** algorithms

... But must be handled carefully

- Wrongful disclosure of data represents a very significant reputational and legal risk
- Data protection regulations are very demanding (GDPR)
- Transition to a cloud-based infrastructure poses further potential security risks

MOTIVATION
RECENT DEVELOPMENTS OF BANCO DE PORTUGAL'S DATA INFRASTRUCTURE

- Integrated Data Management Program
- Developing data culture increases the interest in exploring our data
- Streamlined data processing procedures with centralized data repositories
- BdP DataHub – an effort towards the integration of all data reported to the Bank

- Increasingly sensitive data
(CCR, Banking Deposits Database, Household's Income, …)
- Clear access policy that ensures that people only access sensitive data on a **need-to-know basis…**
- … while making sure that self-service access to data is still possible when appropriate.

MOTIVATION
THE ROLE OF ANONYMIZATION

- **Anonymization techniques** play a vital role in **increasing data protection**, complementing infrastructure security
- Anonymization models are **very diverse**, and their impact on **data integrity** varies
- Building upon previous experiences, BdP has defined a model that **minimizes loss of data integrity** but meaningfully increases **data security**

ANONYMIZATION PRINCIPLES

02

ANONYMIZATION PRINCIPLES
MAIN GOALS

Define

Ensure that core concepts – **personal data, anonymization, pseudonymization** – have **a common definition** at the institutional level

Prototype

Idealize and implement a working prototype for **an internal pseudonymization algorithm**

Idealize

Consider the **key issues** that must be tackled when defining a data privacy model for the **department**, and eventually for **the bank**

Implement

Objectively define the steps that must be taken **to implement the pseudonymization model** and establish the corresponding **governance model**

ANONYMIZATION PRINCIPLES
SOME DEFINITIONS

Personal data

Variables or sets of variables that:

- Individually or when combined;
- Directly or indirectly;

Relate to an individual and allow us to identify them with high confidence.

These variables can be divided into:

- Identifiers
- Quasi-identifiers

Pseudonymization

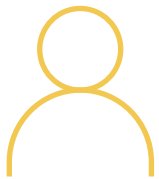
Process that transforms personal data such that the risk of identification is significantly reduced. Methods include:

- Eliminating high risk variables;
- Replacing identifiers with pseudonyms;
- Lowering the level of detail in the data.

There is generally **a strong positive relationship** between the **reduction in risk of identification** and the implicit **loss of information**.

ANONYMIZATION PRINCIPLES
MAIN PIECES

Main information pieces of any data anonymization/pseudonymization model



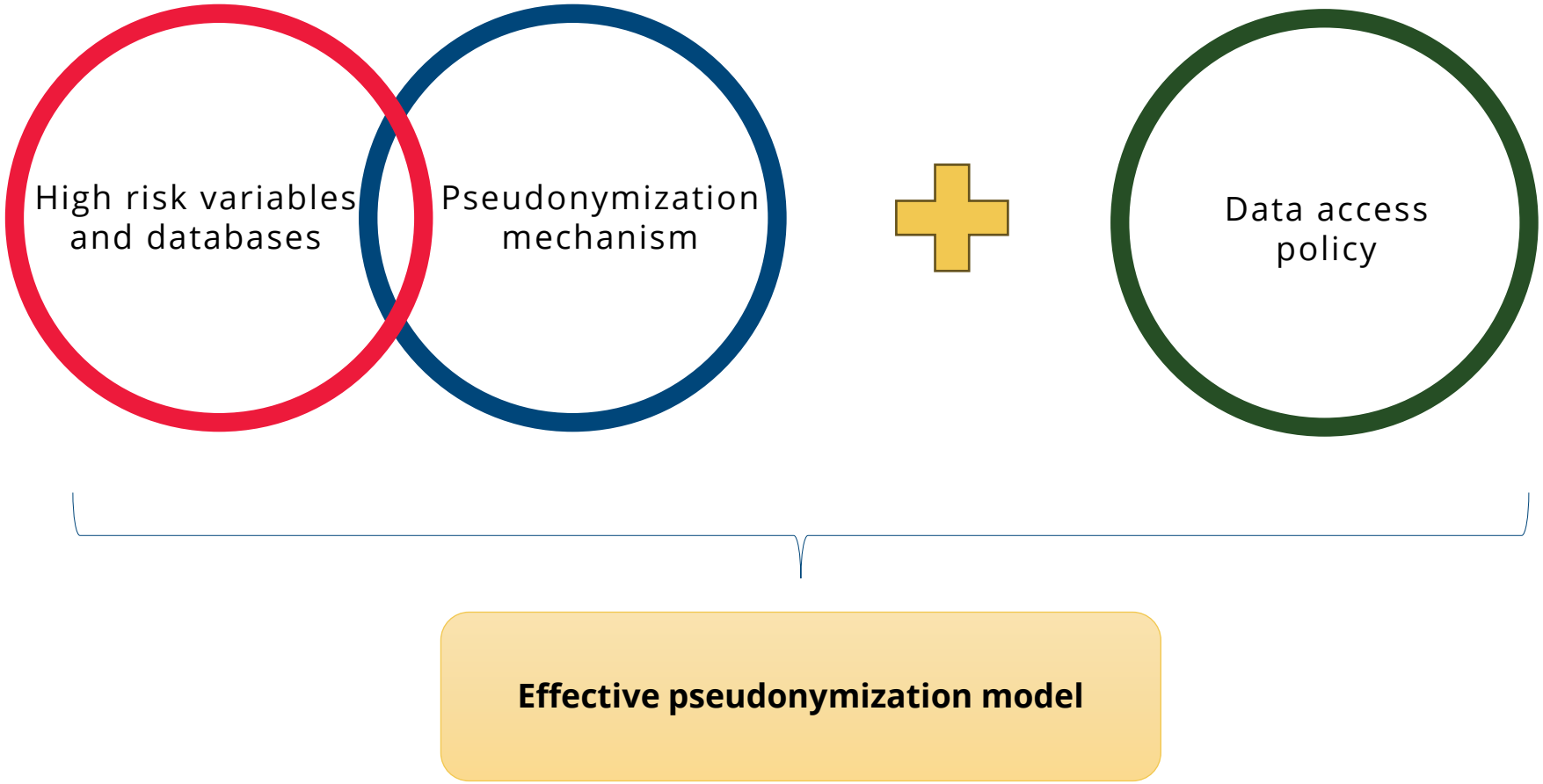
ORIGINAL (IDENTIFIED) DATA

PSEUDONYMIZATION MECHANISM

PSEUDONYMIZED DATA

If a user knows two out of three of these pieces, he can reverse the pseudonymization process and subvert the model

ANONYMIZATION PRINCIPLES
WHAT WE MUST TACKLE



DEFINING THE PATH TOWARDS IMPLEMENTATION

03

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – MAIN TOPICS

1

Critical databases and variables

Evaluate **all databases** where the unit is an individual, and define a list of quasi-identifiers that pose a significant risk of identification

2

Define an access policy

Proposed policy comprised of **3 types of profile**, and focused on **democratizing the access to all pseudonymized data**

Pseudonymization algorithm

Define and implement an **internal pseudonymization algorithm** that can be applied to **all relevant databases** regardless of the types of identifier/quasi-identifiers

3

Data flow

Implement a **data flow** policy that **must be followed** when receiving, using, and sharing **any data where the unit is an individual**

4

DEFINING THE PATH TOWARDS IMPLEMENTATION
PROPOSED MODEL – CRITICAL DATABASES AND VARIABLES

1

We propose:

Evaluate all the databases within the Bank where the unit is an individual (including corporations, when relevant)

Define, at the institutional level, a list of quasi-identifiers that, by themselves or when combined, pose a significant risk of identification

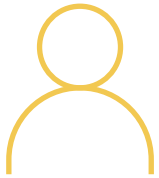
When pseudonymizing a given database, we should refer to this list of variables to determine if any of them should be suppressed

DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – ACCESS POLICY

2

We propose 3 types of profile:



ORIGINAL DATA (TYPE A)

Can access the **minimal necessary set** of **identified data** to carry out his work

Is ignorant of the pseudonymization mechanism and does **not have access to any pseudonymized data**

PSEUDONYMIZATION MECHANISM (TYPE B)

Has access to **all data necessary to conduct the pseudonymization process**

Responsible for implementing and applying the **algorithm**

Restricted to the smallest possible number of people

PSEUDONYMIZED DATA (TYPE C)

Has access to **all pseudonymized databases**

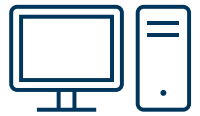
Is ignorant of the pseudonymization mechanism and does **not have access to any identified data**

DEFINING THE PATH TOWARDS IMPLEMENTATION
PROPOSED MODEL – ACCESS POLICY

2

Additionally, we propose:

Type B access should be attributed either to the **IT Department,** in the case of a Bank-wide model, or to **the Information Management Division,** in case of a Department-wide model



Type C access should be given to **every person** that does not have any **Type A** access

Across the Department and the Bank, teams should be set up in such a way as to **isolate all functions that require access to identified data**

DEFINING THE PATH TOWARDS IMPLEMENTATION
PROPOSED MODEL – PSEUDONYMIZATION ALGORITHM

3

We propose a pseudonymization algorithm:

- Recommended and considered sufficient within the context of the GDPR
- Focuses on identifiers, rather than quasi-identifiers
- Preserves the possibility of joining different databases with common identifiers
- Developing an “in-house” model is relatively simple, increasing trust and security
- Ensures reversibility, although the process should only be reversed if strictly necessary

DEFINING THE PATH TOWARDS IMPLEMENTATION
PROPOSED MODEL – PSEUDONYMIZATION ALGORITHM

3

Concerning the algorithm, we propose:

The pseudonymization algorithm **should be defined internally**, should ensure **unique pseudonyms,** and include at least one step that depends on **a key**

The algorithm should be sufficiently **flexible** to be applied to different types of identifiers (numerical, alphanumerical, etc.)

The algorithm should be applied **after defining** the variables that should be made available in the **anonymized database**

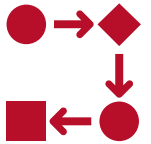
DEFINING THE PATH TOWARDS IMPLEMENTATION

PROPOSED MODEL – DATA FLOW

4

Concerning the data flow, we propose:

Any data sharing, **formal or ad hoc**, must follow the defined **data flows**



The model should be **agile and flexible,** so that people are not tempted to avoid following the data flow to share the data faster

CLOSING REMARKS

04

CLOSING REMARKS
WAY FORWARD

- Thorough **discussion on the viability of our recommendations**

  (particularly regarding the segmentation of different roles and the allocation of the responsibilities defined under our model)

- **Define who's who**

  (assign roles and responsibilities, evaluate the possibility of having a segmentation of roles between people who need to access identified data and those who don't)

- **Experimentation through different use cases**

  (apply pseudonymization techniques to different databases, making them available to specific sets of users)

QUESTIONS

FFONSECA@BPORTUGAL.PT
MFLLOURENCO@BPORTUGAL.PT