Co-funded by the European Union

Project 101146355 – AIML4OS

One-Stop-Shop Artificial Intelligence and Machine Learning for Official Statistics

Project 101146355 – AIML4OS

Co-funded by the European Union

Workpackage 11

ESS AI/ML WORK PACKAGE TITLE

Deliverable	D11.1 – Report describing the training and test sets Rev 1
Month due	May 2025 (for the original version)
Type	R – Document, report
Prepared by	Sónia Quaresma (INE), Alexandre Cunha (INE), João Poças (INE), Przemysław Skrzypczak (Stats Poland), Grzegorz Bujwid (Stats Poland), Marcin Anholzer (Poznan University), Edwin de Jonge (CBS), Gert Buiten (CBS)

Workpackage Leader:

Gert Buiten (CBS)
g.buiten@cbs.nl
CONTACT PHONE – IF APPLICABLE

Deliverable description

This report is produced by Workpackage 11 of the AIML4OS project. The main outcome of this WP are AIML-based models and software that each EU NSI could use to derive a firm-level supply chain network for their national economy, allowing for a range of economic and policy analyses. As such, the WP consists of tasks on

- describing the firm dataset needed for a NSI to derive such a network,
- developing and training the Machine Learning models on an actual national firm-level network dataset to be used by a NSI,
- and the ask to derive a national firm-level network dataset that can be used as training set for the ML models.

This report is the revised version of Deliverable D11.1 of WP11 and documents the successful completion of Task 11.1 "Create one or more training sets with user/supplier links". It is the revised version of the original deliverable that was accepted by Eurostat as good enough, but with a number of suggestions for further improvement. These suggestions have been implemented in this current version.

In this task, the following steps were undertaken:

1. the available input data for the firm dataset as well as the input link data for creating the national firm-level network dataset have been identified, showing that the necessary data is available for Portugal. The firm data consist of generally available characteristics and variables for all firms such as NACE class, size class, geographic location, turnover.
2. Choose the best method to derive a prototype network dataset from the available input link data, building on methods developed in at the Central Bank of Belgium and several other countries. Special attention was given to data cleaning and anomaly techniques to pre-process the raw data. After that, a firm-level network dataset was derived from the Portuguese input datasets. The network dataset was enriched with the variables from the firm dataset.
3. Finally, the resulting data was converted and expanded into a set of interlinked datasets that are the input for the next task of WP11: T11.2 "Use ML to develop models for deriving binary networks from firm-level data".

Abstract

This report is being submitted in accordance with deliverable 11.1 within workpackage 11. The purpose of this deliverable is to document the process of creating the training dataset for a machine learning model to reconstruct buyer–supplier networks among enterprises, using administrative firm data and transactional links from electronic tax invoices. This integration enables a coherent, rich dataset for network structure prediction. The quality of the resulting dataset has been validated by comparison of network metrics with other, similar datasets in the literature.

1. Introduction

In the rapidly evolving field of data science and machine learning, reconstructing buyer–supplier networks among enterprises offers significant value for economic and policy analysis. This document outlines the creation of training and test datasets to support the development of a machine learning model aimed at predicting such networks, within the scope of the AIML4OS initiative. Enterprises, as the network's nodes, are characterised using information originating from firm-level administrative data; already processed and integrated as will be explained — including firm size, sector of activity, and geographical location. The links between them, indicating economic activity, are inferred from transactional data derived from electronic tax invoices, representing the exchange of goods and services

By integrating these two complementary data sources for Portugal, the firms dataset and the network dataset, we aim to produce a coherent and comprehensive dataset capable of supporting the training of robust predictive ML models which would be used for predicting network datasets for other NSI's. The following sections describe:

- the methodology used for compiling and preprocessing the network dataset as a training dataset;
- the creation of a synthetic dataset to enable secure collaboration; and
- the key challenges faced in ensuring the dataset's accuracy and representativeness.

This structured approach lays the groundwork for advancing our understanding of enterprise interactions and fostering reliable model development.

2. Creating a network dataset

2.1. Deriving a firm dataset

The firms form the nodes in the buyer/supplier network. The overall assumption is that the characteristics of two firms is predictive for them having a buyer-supplier link. The firm dataset holds detailed information on enterprise characteristics, including size, sector of activity, legal form, and location, — thereby forming the basis for characterising the nodes of the network.

The firm dataset is an essential ingredient for the ML models: they are the essential "predictors" of the models. Because machine learning models are inherently shaped by the structure and content of the data they are trained on, any National Statistical Institute (NSI) wishing to apply the WP11 models must first construct a firm dataset with an equivalent structure. Without this crucial step, the models cannot be reliably transferred or produce meaningful results. The selection of variables should be readily available for EU NSI's. The variables of the firm dataset identified are in Table 1. The variables in bold should be available for all NSI's. The other ones are included for the trainingset for WP11 in order to be able to assess their predictive power.

Table 1: variables in the firm dataset

YYYY	**Reference year**
ID	**Business identification**
NACE_SEC	**NACE sector**
NACE	**NACE cod**
NUTS3	**Nomenclature of territorial units for statistics (small regions)**
TO	**Turnover**
PURCH	Purchases (goods and services)
NPE	**Number of Persons Employed**
ACT	Actives (Net Assets)
WAGES	Expenditure on personnel
DIM	Company dimension based on Number of Persons Employed, TurnOver and Actives
FTA	Fixed tangible assets
FTI	Fixed tangible actives' investments
IAG	Intangible assets, without goodwill
IAI	Intangible assets' investments

2.2. Creating a buyer/supplier network dataset

To construct a robust buyer–supplier network dataset for Portugal, several administrative data sources were initially considered and evaluated based on their relevance, completeness, and suitability for the intended analytical framework. Following this assessment, two key elements were selected as both sufficient and appropriate for the purpose of having the nodes and the links.

As discussed above, the network dataset has two key elements. The first is the set of network nodes, which comes from the dataset with characteristics and variables on individual firms. The second element is the set of edges c.q. links between the nodes, which comes from the dataset on buyer-supplier relationships between firms. These relationships are based on the tax invoice records, which make it possible to identify supplier–buyer relationships in the exchange of goods and services.

In this section we describe some of the characteristics of these two key elements, including how they may be created from administrative data sources, and how they can be integrated and validated[1]. It further describes the methodology within this specific Workpackage 11 of AIML4OS adopted for data integration and variable selection and introduces the resulting dataset. A more detailed characterisation of the inter-firm connections — including the directionality and intensity (weight) of the relationships — will be provided in the following section.

2.2.1. Selected firms and network input dataset and data preparation steps

In early 2006, as part of a government-led initiative to simplify and modernise public administration — with the goal of making life easier for both enterprises and citizens — several national public entities, including Statistics Portugal, collaborated in the development of an integrated system for administrative data collection. This created the possibility for developing a firm-level supply chain network dataset, but also is important input for the firms dataset itself.

This initiative enabled companies to fulfil multiple legal obligations through a single reporting process, namely:

- Submission of annual accounting and tax statements to the Ministry of Finance (Tax Administration).
- Registration of accounts in accordance with commercial registry legislation.
- Provision of statistical business information to Statistics Portugal, in compliance with requirements of the European Statistical System.
- Reporting of accounting data to the Portuguese Central Bank, in line with its responsibilities within the European System of Central Banks.

This integrated system — known as the **Simplified Business Information (IES)** — was officially established by **Decree-Law No. 8/2007 of 17 January**. The implementation of IES was only possible thanks to the strong collaboration between all stakeholders. In addition to the participating public institutions, key support came from the Chamber of Chartered Accountants, software providers to enterprises, and financial and insurance supervisory authorities. Nonetheless, the crucial enabling factor was the solid policy support that sustained the system's creation.

[1] The approach we follow is based on experiences in other countries. See Magerman, G., Dhyne, E. and Rubínová, S., (2015). *The Belgian production network 2002-2012*, NBB working paper research nr. 288; Mungo, L, A. Brintrup, D. Garlaschelli and F. Lafond, *Reconstructing supply networks*, arXiv:2310.00446v1 [econ.GN] 30 Sep 2023; Borsos, A., M. Stancsics, *Unfolding the hidden structure of the Hungarian multilayer firm network*, MNB Occasional Papers, No. 139; Criscuolo, C., A. Dechezleprêtre, L. Guillouet, G. Lalanne, F. Manaresi, *Estonia's firm-level production network: Lessons for industrial policy*, OECD Science, Technology and Industry Working Papers 2024/13 and
Alfaro Ureña, A, Mariany Fuentes Fuentes, Isabela Manelici, José P. Vásquez, *Costa Rican Production Network: Stylized Facts*, Documento de investigación N. 002, 2018, Banco Central de Costa Rica

From the outset, Statistics Portugal defined a clear strategic objective for its participation: the gradual replacement of the **Annual Business Survey** with statistics derived exclusively from administrative data. Remarkably, this objective was achieved in the first year of IES implementation, without disruptions in the statistical series. Despite the methodological differences, the data submitted, with the IES as the main source, for companies, were broadly consistent with those collected in previous years via survey.

Prior to IES, the Structural Business Statistics (SBS) relied on the sample annual business survey. With the advent of IES, this survey was fully discontinued, and all overlapping variables were removed from other data collections, leading to a significant transformation of the production system.

The **Integrated Business Accounts System (SCIE)** is the foundational framework to produce structural business statistics. It is an integrated system designed to cross-check, compare, and analyse data from multiple sources related to business statistics — with the **Simplified Business Information (IES)** serving as the principal input. The IES is complemented, with data for individual enterprises (sole proprietors and self-employed workers), received via the protocol established between INE and the Tax Authority.

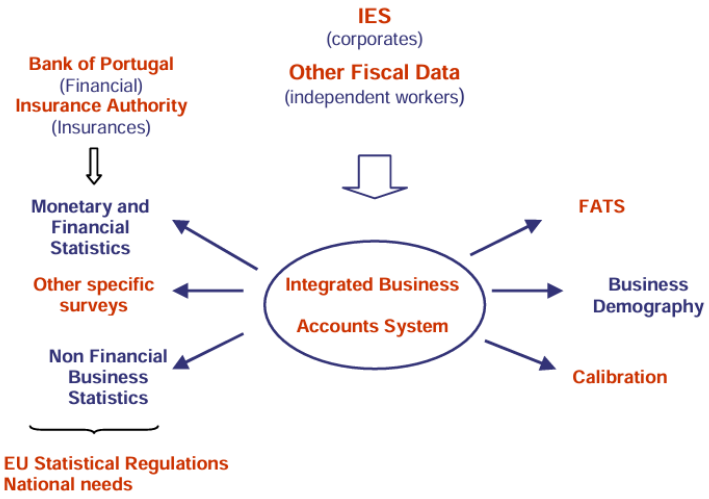

Figure 1 – Structural Business Statistics based on IES

To ensure the completeness of the dataset, SCIE performs a systematic comparison with other data sources such as the Business Register. This comparison enables the accurate determination of the population of active enterprises to be considered each year.

For units that do not respond — whether due to delays or total non-response — an estimation procedure is applied. This method relies on existing responses (primarily from IES) and is conducted by strata, allowing the imputation of values for all required variables. Through this process, it becomes possible to obtain complete and coherent data for the entire enterprise population, even in the presence of partial or missing responses.

The consolidated SCIE register thus contains a wide range of variables designed to describe various characteristics of enterprises. However, given the extensive number of available variables, only a carefully selected subset was chosen for inclusion in the current analysis.[2]

[2] Here, we build on previous work on applying ML-approaches to network reconstruction, e.g. by Mungo, L., F. Lafond, P. Astudillo-Estévez and J.D. Farmer, *Reconstructing production networks using machine learning*, January 2022, INET Oxford Working Paper No. 2022-02 and Buiten, G., E. de Jonge, J. Vuik, J. Meijers, G.

This reduction was guided by two main considerations:

1. Model Training Requirements:
   The selected variables in the firms dataset were intended to form the basis of a training dataset used to develop a predictive model for identifying the existence (or absence) of a link between two enterprises. To ensure that this model can be applied consistently, we must be able to retrieve the same set of characteristics for any other enterprise to which the model is applied. Using too many or overly specific variables could compromise generalizability and the ability to replicate the feature set across the population, or in different countries.

2. Data Availability and Consistency:
   Some of the variables available in SCIE are conditional on the enterprise's economic activity (e.g., only relevant for industrial or service-oriented firms) and are therefore not consistently reported across all sectors. Including such variables could introduce bias or missing data problems, limiting the model's robustness and interpretability.

As a result, the variable selection process prioritized universally available, economically meaningful, and non-sector-specific indicators that allow for both reliable model training and scalable prediction across the enterprise population — thereby forming the basis for characterising the nodes of the network.

The network data source consists of electronic tax invoice records (e-Fatura), which allow for the direct identification of buyer–supplier relationships between firms through the exchange of goods and services. In Portugal, the electronic transmission of invoices issued by individuals or legal entities with a registered office or permanent establishment in national territory to the **Tax and Customs Authority (AT)** is mandatory. This obligation was introduced as part of a broader strategy of administrative simplification and anti-fraud measures implemented by the Tax Administration.

Within the framework of a cooperation protocol between Statistics Portugal (INE) and the Tax and Customs Authority, monthly invoice data collected through the e-Fatura system is transmitted to Statistics Portugal.

This administrative dataset encompasses all invoicing activity reported electronically by the issuer, regardless of whether the buyer has formally requested an invoice.

Two major challenges arise from the use of this data source:

1. Ensuring regular and timely monthly transmission from the data provider.

2. Managing the extremely large volume of records, which currently exceeds 100 million invoices per month.

To ensure data quality and usability, several data treatment procedures are applied[3]:

Mooijen, S. Hooijmaaijers, P. Bogaart, *Reconstruction method for the Dutch interfirm network including a breakdown by commodity for 2019/2021*. Version 1.1, August 2024

[3] Here we follow the approach used by the Belgian Central bank for deriving a firm-level supply chain network based on more or less similar administrative data, as described in: Emmanuel Dhyne, Glenn Magerman and Stela Rubínová, *The Belgian production network 2002-2012,* Working paper research oct 2015 nr 288, Belgian Central Bank

- Outlier Detection and Imputation:
  Extreme values are flagged and imputed, particularly when positive taxable amounts equal or exceed €100 million and deviate more than three standard deviations from the sectoral average. For negative taxable values, those with the highest magnitude are analysed case by case.

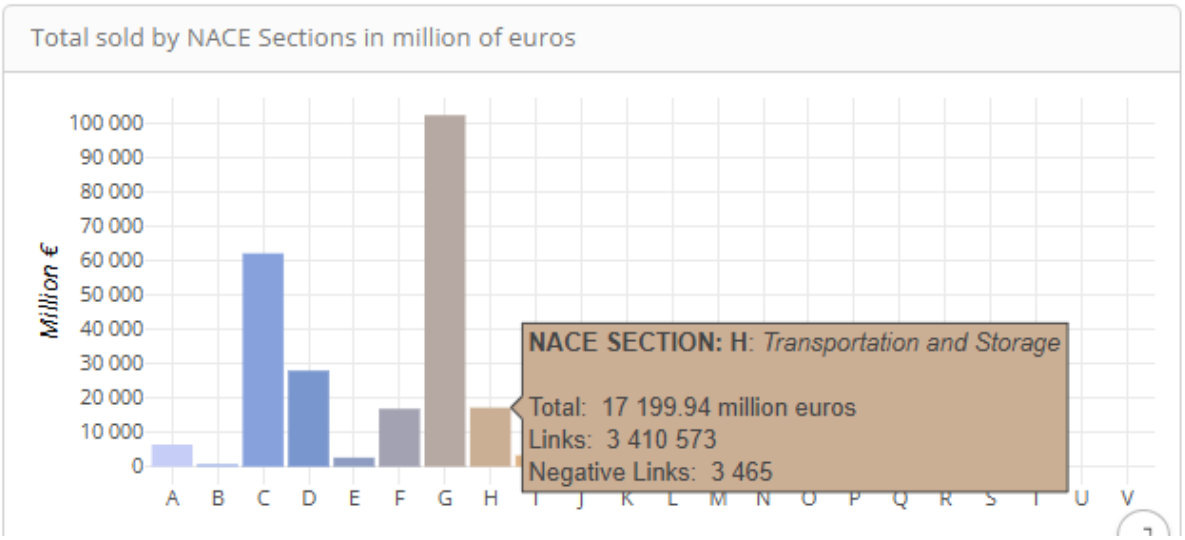

Figure 2 - Total Sold by NACE Sectors in million of Euros in 2022 (the pop-up illustrates the existence of negative inks)

- Correction of Negative Values:
  Negative taxable values exceeding €100,000 in absolute terms are corrected when a corresponding symmetrical value (95–100% similarity) is identified in the previous four months. These negative values typically result from corrections to previous erroneous submissions. The total taxable value of the involved transactions remains constant; only the temporal allocation is adjusted. In 2022 only 86 272 (0.21%) links with negative value remain after this treatment.

- Imputation of Missing Values for Key Enterprises:
  For a small but significant subset of enterprises — selected based on turnover and number of employees — missing data is identified and imputed using historical behaviour and time series patterns.

A detailed list of the variables used from the e-Fatura dataset can be found in Annex 1. Please note that the derivation of the firm-level production network is not a regular task at Statistics Portugal. It was carried out specifically for the AIML4OS project as a one-off activity, given that the process is highly time-consuming and resource-intensive. Future updates of the network dataset are therefore not currently envisaged.

The basic general workflow looks like this:

2.2.2. Description of the methodology for processing geographical information

The reference year 2022 was selected primarily due to the characteristics of the SCIE data. The administrative data from IES is not used directly but goes through complex processing and validation steps. Given the inherent delay of approximately two years in the processes of data transmission, treatment, and validation, choosing 2022 was necessary to ensure that the data would be available and ready for analysis starting in April 2024.

The firms dataset did not require significant transformation, as it had already undergone extensive processing and validation steps to be integrated into SCIE. This prior harmonization ensured consistency and reduced the need for further preprocessing.

Most of the data handling efforts were therefore concentrated on the E-Fatura dataset. These data, originally recorded at a monthly frequency, were aggregated to the annual level to match the granularity of the SCIE data. The selection of 2022 as the reference year in both sources also ensured temporal alignment, allowing for coherent integration and comparative analysis across datasets.

YYYY *Int64*	SUPPLIER *Utf8*	BUYER *Utf8*
2022	63cda26405dccbe50767d93038cdff5c	ede6e20fdeec34c7f518ff8d1d86d41a
2022	63cda26405dccbe50767d93038cdff5c	4629303b97791a7f73a26dd7c12234b0
2022	63cda26405dccbe50767d93038cdff5c	cdef431 2ec6ad168f0d2cf661c3be406
2022	63cda26405dccbe50767d93038cdff5c	04070e5594eadcb52c5b535327ecc557
2022	63cda26405dccbe50767d93038cdff5c	a4cac0e947f0cece84c84d83e612f27b
2022	63cda26405dccbe50767d93038cdff5c	5ed9b3b3ec7972a88e7e25e5d2451ff0

Figure 3 - part of E-Fatura dataset

An important predictor for a supplier-buyer link is geographical distance between two firms. This variable is not directly available but can be derived using the firms dataset. The geographic distance between supplier and buyer firms was calculated for the existing buyer supplier links based on the location information available. However, two key limitations emerged in this process. Firstly, the address information corresponds exclusively to the headquarters of each firm, and not to the specific establishment involved in the transaction. This means that the calculated distances may not accurately reflect the true origin and destination of goods or services, especially in cases where firms operate multiple branches or facilities.

Secondly, the addresses themselves are represented only by municipality codes, which provide a relatively coarse spatial resolution. This further limits the precision of the distance calculation, as it assumes a single central point per municipality, regardless of the actual intra-municipality variation in firm locations.

As a result, the aggregated data may include transactions originating from different establishments within the same firm, or even between a head office and a separate branch. In some edge cases, the headquarters may not be involved in commercial activity at all, serving solely administrative functions. These limitations could potentially introduce noise into the model and affect its predictive

performance. While it was not possible to correct for this issue at the current stage, it has been clearly identified and documented to inform future refinement and interpretation of the model outcomes.

The final challenge was of a more practical nature: how to calculate the distances between municipalities, given that this was the only geographic information available. To address this, two alternative methods were devised. The first involved calculating the distance by road between municipalities, taking into account the actual road network. The second approach computed the straight-line (geodesic) distance between the centroids of each municipality.

Both methods were implemented using libraries from the R programming language. For the road-based distances, the osrm::osrmRoute function was used, leveraging OpenStreetMap routing services to reflect realistic travel paths. For the centroid-based approach, the raster::pointDistance function was employed to measure the geographical distance between two points on a plane. These complementary methods aimed to provide a more comprehensive estimation of spatial proximity between firms, acknowledging the limitations of the available address data and finally the smaller distance is chosen. With the current firm dataset, the centroid-based distance is always shorter. This is due to the fact that we currently only have the municipality centroid and not the exact geographical location of the individual firm. The software is in place for the case that more detailed firm coordinates become available.

When the two municipalities are the same, the reported value corresponds to the median distance between the centroid and the municipality's boundary points. The function also allows for alternative calculations, such as the average distance to the boundary points, the distance to the furthest point, or the distance to the closest point on the border. However, after discussion and visual inspection of several cases, the median distance was considered the most appropriate and was therefore adopted.

Finally, there are no thresholds to the reporting through e-fatura and all transactions albeit aggregated are considered.

3. Synthetic dataset for development and testing purposes

3.1. Synthetic Dataset Construction: the approach

Since the original dataset cannot leave the premises of Statistics Portugal due to confidentiality constraints, having a synthetic version is essential to enable collaboration. Without it, external partners would be unable to explore the data, experiment, or begin developing modules for machine learning training and testing.

Using the {synthpop} R-package[4], the firm dataset was synthetically generated, allowing them to be shared externally without compromising the confidentiality of the original data.

Because these synthetic datasets are subject to fewer restrictions regarding storage and handling, they are made available as part of Work Package 11 (WP11), specifically within the IT environment provided by Onyxia (WP3). In summary, for a set of p variables, the data synthesis process follows these steps:

[4] https://www.synthpop.org.uk/

1. Draw a simple random sample from the first variable x_1^{obs} and define it as x_1^{syn};
2. Define a model $f(x_2^{obs} \mid x_1^{obs})$, and draw x_2^{syn} from $f(x_2^{syn} \mid x_1^{syn})$;
3. Define a model $f(x_3^{obs} \mid x_1^{obs}, x_2^{obs})$, and draw x_3^{syn} from $f(x_3^{syn} \mid x_1^{syn}, x_2^{syn})$;
4. Continue this process successively until reaching $f(x_p^{syn} \mid x_1^{syn}, x_2^{syn}, \ldots, x_{p-1}^{syn})$.

This method ensures the synthetic data preserves the relationships between variables while protecting the confidentiality of the original dataset.

With the intention of making the synthetic dataset also smaller and more manageable than the real files only a subset of enterprises was selected for this exercise. The base dataset used with the {synthpop} package was constructed by identifying relevant fiscal numbers (NIFs) that refer exclusively to legal entities ("pessoas coletivas"), excluding both individuals ("pessoas singulares") and sole proprietors ("empresários em nome individual"). The selection of these relevant NIFs was based on four specific criteria:

- STA Criterion: Based on the Statistical Business Register (FUE), firms were selected if they had a Gross Added Value (VVN) over €50 million, or VVN above €10 million and more than 200 employees (NPS), and an economic activity complexity (STA) below 30.
- DEE Criterion: Based on the Integrated Business Accounts System (SCIE), firms were included if they had a VVN over €50 million and Net Assets over €43 million, or if they employed more than 250 people.
- peso10 Criterion: Based on electronic invoice (e-Fatura) data, this criterion selected firms whose annual taxable turnover (VT) exceeded 10% of the total VT for their respective 3-digit CAE (economic activity classification).
    - NIFs belonging to certain CAEs were excluded: CAEs like '061', '071', '091', etc., were excluded due to fewer than 100 records per year. Others like '182', '192', '202', etc., were excluded because in some years, a single NIF accounted for more than 90% of the total turnover for that CAE, potentially distorting representativeness.
- DCN Criterion: A set of NIFs was directly indicated by the National Accounts Department (DCN) of Statistics Portugal as relevant for inclusion, based on their importance to national accounts estimations.

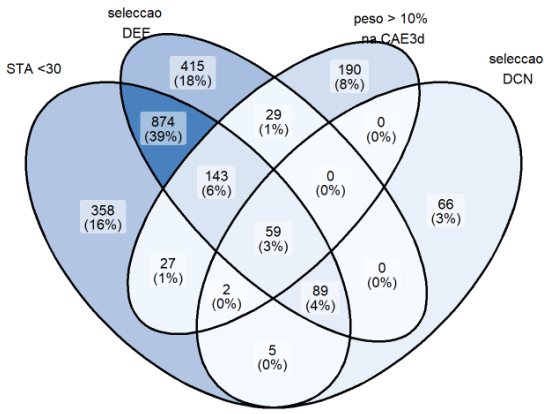

Figure 4 - distribution of the 2040 enterprises according to all the Criteria

This rigorous multi-source selection ensured that the synthetic data reflected a representative and analytically valuable subset of economically significant firms, making it especially suitable for model training and development while respecting data confidentiality constraints.

A total of 2,040 enterprises were selected for the synthetic dataset exercise. These firms are primarily concentrated in the major regions of Lisbon and Porto, reflecting the geographic distribution of economically significant business activity in the country.

Figure 5 - National distribution of the companies

An analysis was conducted to gain a clearer understanding of the distribution of company sizes within the selected sample used to generate the synthetic dataset. Although it was already anticipated—based on the previously described selection criteria—that the process would favour the inclusion of medium and large enterprises, this analysis aimed to quantify that bias. By examining the proportions by company size, we were able to confirm and better characterise the expected concentration in larger firms, thereby reinforcing the representativeness of the synthetic dataset for this specific segment of the business population.

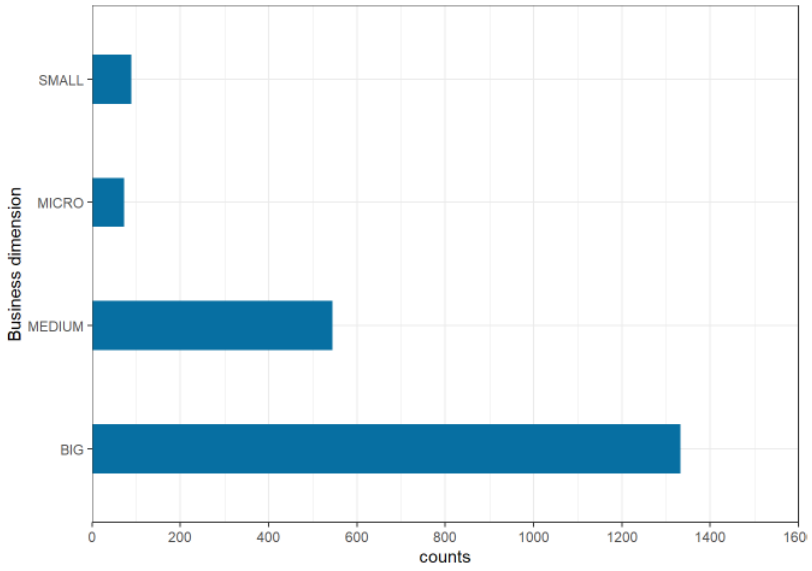

Figure 6 - Enterprises by size on the synthetic dataset

3.2. Synthetic Dataset Construction: Description of the results

It was clear from the outset that the size distribution achieved on the synthetic dataset would not reflect the true composition of the enterprise population in the real dataset. In reality, the structure is expected to be significantly skewed towards smaller firms. This is particularly due to the absence of thresholds on transaction reporting, which results in a much broader inclusion of micro and small enterprises in the administrative records. Verifying this discrepancy in size distribution was therefore one of our initial concerns, as it directly impacts the generalizability of any exploratory analyses or machine learning models developed using the synthetic dataset. Understanding this divergence is essential to ensure that users of the synthetic data remain aware of its limitations and the specific context in which it was generated.

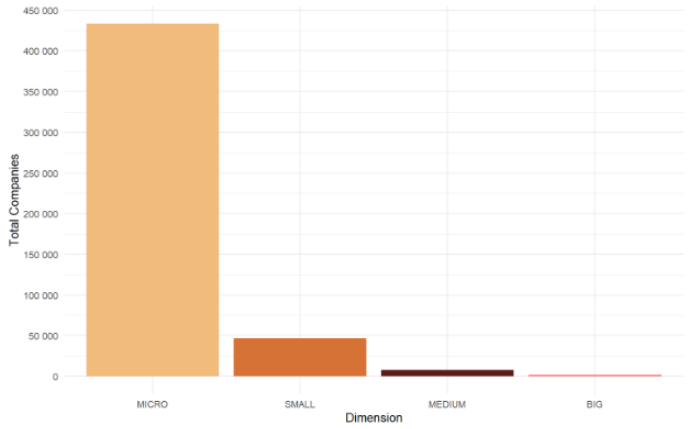

Figure 7 - enterprises by size on the real dataset

It became evident that the synthetic dataset was not suitable for training a machine learning model, but rather served the more limited purpose of supporting general input validation during software development. Given that the real dataset cannot be made publicly available, it was crucial to understand the distribution of features as this will significantly influence the model.

Descriptive Statistics

Variable	Type	Distinct Values	Null Values	Most Common Value	Min	Mean	Median	Max	Lower Percentile *10%*	Higher Percentile *90%*
YYYY	Character	1	0	2022						
ID	Character	489 351	0							
NACE_SEC	Character	18	0	G						
NACE_COD	Character	813	0	68100						
NUTS3	Character	27	2 254	1A0						
TO	Numeric		0		0	1 081 616	80 534	13 448 547 665	0	948 530
PURCH	Numeric		0		-12 880 141	843 404	47 588	13 558 056 568	2 071	681 969
NPE	Numeric		0		1	7	2	26 859	1	10
ACT	Numeric		0		0	3 101 445	115 910	197 735 778 874	7 505	1 509 308
WAGES	Numeric		0		0	121 401	16 271	385 041 563	0	151 742
DIM	Character	4	0	MICRO						
FTA	Numeric		8 553		0	317 335	8 000	3 592 440 000	0	304 508
FTI	Numeric		8 553		0	57 632	0	372 300 918	0	48 000
IAG	Numeric		8 553		-122 591	71 423	0	2 526 750 000	0	267
IAI	Numeric		8 553		-3 250 304	9 050	0	307 747 000	0	0

Figure 8 - table with the descriptive statistics of the Firm Characteristics Variables

Accounting for the drawbacks mentioned, the synthetic dataset does serve important goals in the project. It allows for software development, since the format is identical to the real dataset. Also, it allows for software testing and debugging, since it also contains all the relevant characteristics of the real dataset (e.g. in terms of NACE codes, size classes etcetera). Code that has been developed, tested and debugged on the synthetic dataset can be downloaded by users and run on their own datasets to produce results. This enables an approach where developers and trainers can work from other countries as well.

Synthetic datasets for other purpose can be created once the reconstruction models are good enough. If the reconstruction model is applied on a dataset for a synthetic population of firms and the necessary characteristics and variables, a synthetic firm-level network dataset can be created that has no resemblance with any real country. By adding some random noise to that dataset will create a dataset that can be used for e.g. educational purposes or even certain types of policy research.

4. Validation of the network dataset

To validate the created network dataset, an internal report[5] was produced in which the most relevant network metrics were reproduced an analyzed the methodology outlined in the study by Bacilieri et al[6]. That study contains the network metrics for network datasets based on large administrative datasets for a number of other countries that also have such datasets. It shows above all that there are strong similarities across countries in the structure and characteristics of supply chain networks. Reproducing these metrics for our own network dataset allows a comparison, that shows that our dataset has the same characteristics as other network datasets. This comparison thus shows that the Portuguese dataset has the right quality for being used for training network reconstruction models and that those models can be applied in other countries as well.

The comparison was done for the following metrics: network degree, the power-law fit of the degree distributions, the degree correlation, the degree assortativity, reciprocity, clustering, average shortest path lengths, strength of the weights, the strength – degree relationship and the influence vector. The metrics were calculated not only for the network as such but also for breakdowns by firm size, industry and NUTS-regions. And as mentioned, the results for Portugal are similar to those for the networks analyzed by Bacilieri et al.

Below, some examples of the structure of the Portuguese network are given. First of all, the average shortest path length by firm size. This is a measure for how well the nodes in the network are connected to other nodes. As expected, the average path length is shortest for the largest firms and longest for the smallest ones. Most of the large firms are only two steps away from any other firm in the network. For the smallest firms, the median distance is eight steps but some are even fourteen steps away from other firms in the network.

[5] Alexandre Cunha, *Estimating Firm-level Supply Chain Networks,* 2025-09-25

[6] Bacilieri, A., A. Borsos, P. Astudillo-Estevez and F. Lafond, *Firm-level production networks: what do we (really) know?*, May 2023, INET Oxford Working Paper No. 2023-08

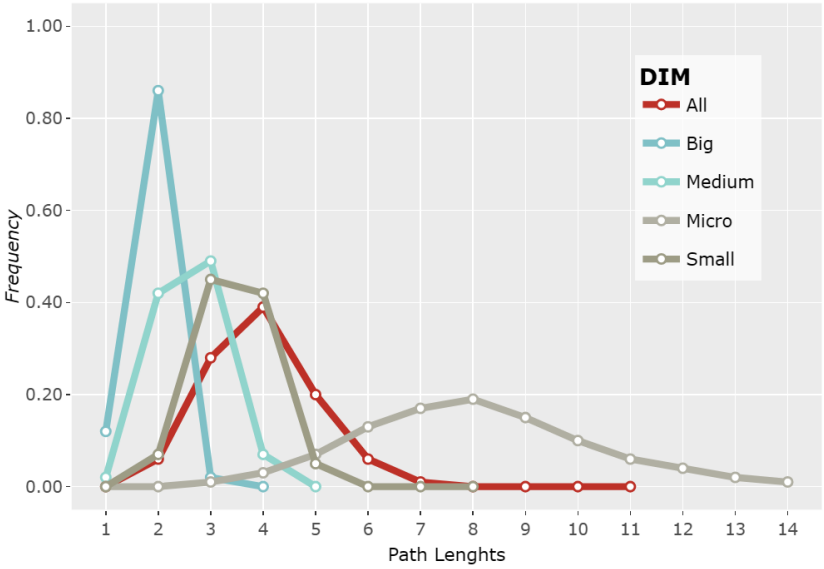

Figure 9 Average shortest path length by firm size

Another metric is the distribution of the buyer/supplier links. In figure 10, the percentage distribution is shown for the number of links between firms by size class. Almost 22% of the links is from micro firms to other micro firms. Only 2% of the links is from big firms to micro firms – whereas just over 15% is from small to big firms. Please note that the percentages are calculated on the number of links and not the weight of the links.

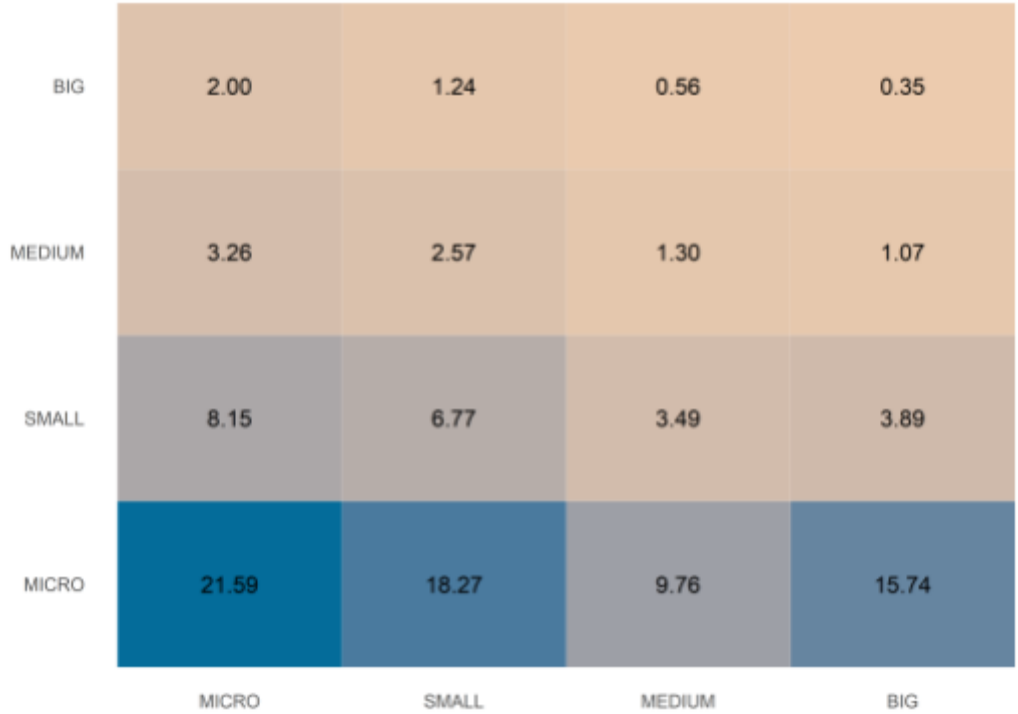

Figure 10 Distribution of the number of buyer/supplier links by size class

For a deeper analysis of the degree distribution, we also computed the empirical Complementary Cumulative Distribution Function (CCDF) separately for in- and out-degrees. As expected, and in line with the findings of the Bacilieri study, this dataset also displays very broad degree distributions with

heavy tails. Also consistent with that paper, we observe a pronounced asymmetry: the maximum number of suppliers (in-degree) is smaller than the maximum number of customers (out-degree).

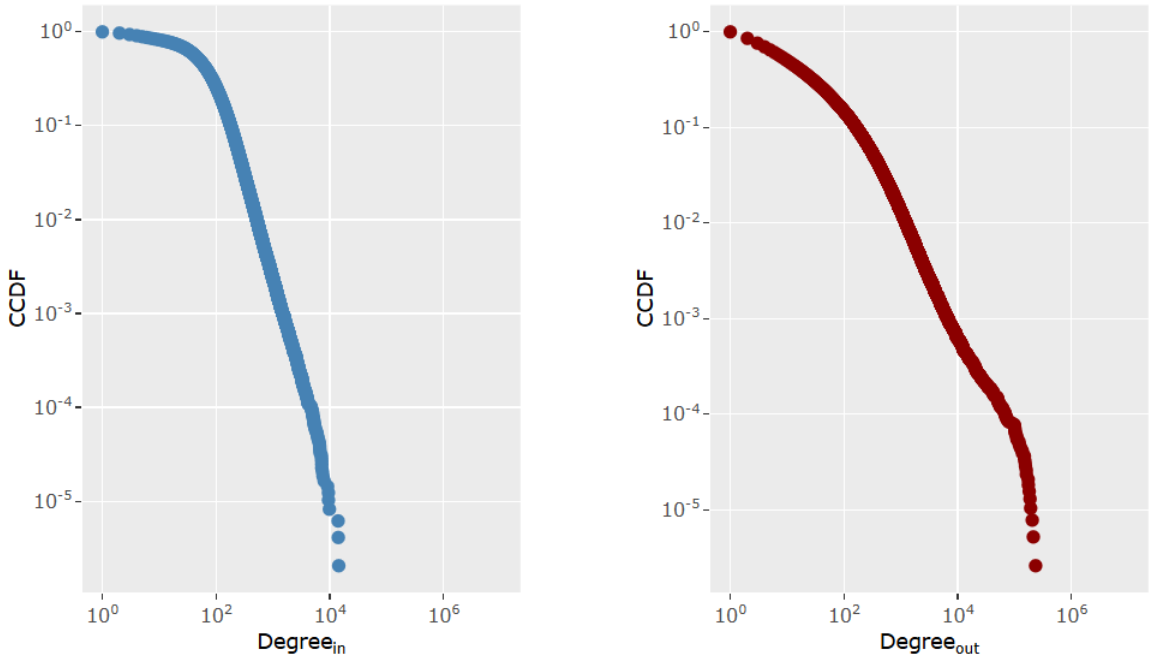

Figure 11 Distribution of the in and out degrees

5. Analysis for preparing modeling

5.1. Calculating link characteristics

In the buyer–supplier network the links were produced with the administrative dataset that encompasses all invoicing activity reported electronically by the issuer.

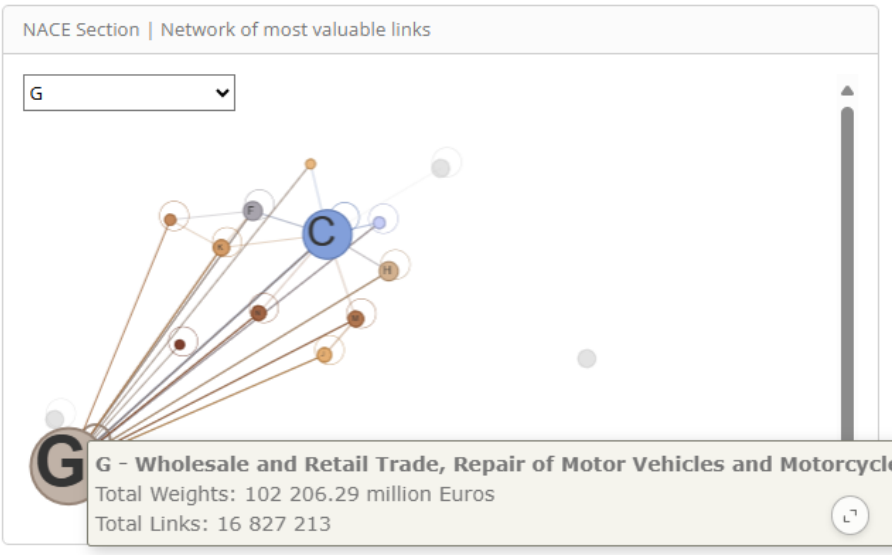

Figure 12 – example of the buyer-supplier network

To better understand the distribution and characteristics of these links, a study was conducted using the NACE classification — the Statistical Classification of Economic Activities in the European Community. NACE categorizes enterprises according to their primary economic activity, allowing for the identification of sectoral patterns in trade and transactional behaviour. This categorization is crucial for machine learning models, as sector-specific behaviours and relationships can influence the probability and nature of buyer–supplier connections. For instance, some sectors may have a higher tendency to cluster, exhibit vertical integration, or maintain stable supplier networks, all of which can affect the predictive performance and structure of the model. Understanding these dynamics ensures the model accounts for sectoral heterogeneity and avoids biases that could arise from overly generalised assumptions.

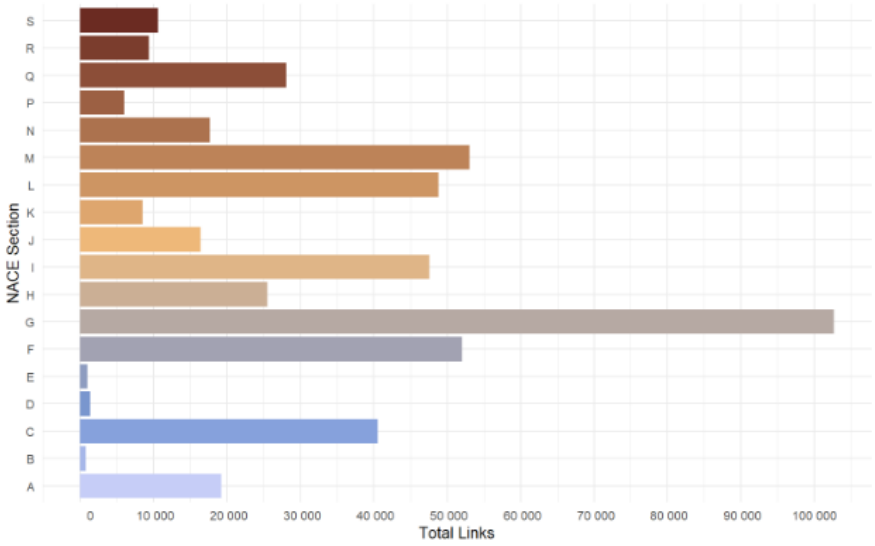

Figure 13 - Links by NACE sector

In particular, it was observed and documented that for certain sectors — especially at the more granular levels of the NACE classification hierarchy (division) — no links were present in the training data. This absence implies that the model will lack exposure to representative examples from these divisions, making it impossible to learn meaningful patterns for them. Consequently, this limitation must be acknowledged, as it may significantly hinder the model's ability to generalise or make accurate predictions in these underrepresented areas.

NACE	Description	% Missing
Sector B	B - Mining and quarrying	27,27
Division 05	05 - Mining of coal and lignite	100
Division 06	06 - Extraction of crude petroleum and natural gas	100
Division 07	07 - Mining of metal ores	33,33
Sector O	O - Administrative and support service activities	100
Division 84	84 - Public administration and defence; compulsory social security	100
Sector T	T - Activities of households as employers and undifferentiated goods- and service-producing activities of households for own use	100
Divisions 97 & 98	97 - Activities of households as employers of domestic personnel 98 - Undifferentiated goods-and services-producing activities of private households for own use	100
Sector U	U - Activities of extraterritorial organisations and bodies	100
Division 99	99 - Activities of extraterritorial organisations and bodies	100

For a complete list by NACE and division check the Supporting Information section.

It is equally important to ensure a balanced representation of links across all geographical regions to enable the model to generalize effectively. A dataset heavily concentrated in a few regions may lead

the model to learn patterns specific to those areas, which can hinder its performance when applied to less represented regions. In this context, particular attention was given to identifying regions that might disproportionately influence the model's learning process — notably the Greater Lisbon and Greater Oporto areas, which dominate in terms of both the number of enterprises and transactional volume.

At first glance, this regional imbalance can introduce bias into the model, potentially leading it to overfit to the economic dynamics of these two major metropolitan areas while underperforming in others. The network science literature shows that the geographical distribution of economic activity in cities and other areas usually follows a power law scaling (see e.g. Geoffrey West, *Scale The Universal Laws of Life, Growth, and Death in Organisms, Cities, and Companies*, 2018). In fact, the majority of connections are formed locally: the median distance between a firm and a supplier is 30 km in Japan (Bernard, Moxnes and Saito, 2019 and 20 km in Belgium (Dhyne and Duprez, 2017).[7] This means that with the right kind of ML modeling techniques this should not lead to a bias. It is above all important to use normalization of geographical distance either in the models themselves or as a separate data preparation step.

Monitoring the distribution of links across regions remains however important, not only for fairness and accuracy but also for understanding the model's limitations and areas requiring further refinement. The real dataset exhibits the same regional patterns previously observed in the synthetic dataset.

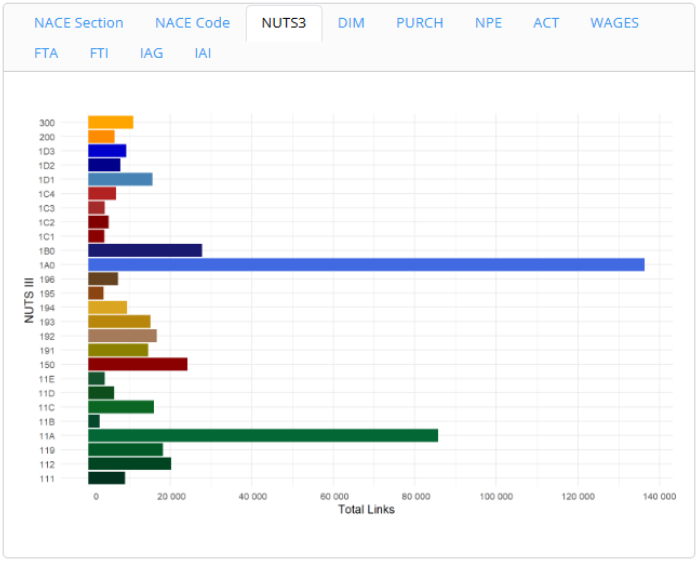

Figure 14 - links distribution by nuts3 region

[7] See also Bacilieri, A., A. Borsos, P. Astudillo-Estevez and F. Lafond, *Firm-level production networks: what do we (really) know?*, May 2023, INET Oxford Working Paper No. 2023-08; Bernard, A., A. Moxnes and Y. Saito (2019), *Production networks, geography, and firm performance*, Journal of Political Economy, Vol. 127/2 and Dhyne, E. and C. Duprez (2017), *It's a Small, Small World... A Guided Tour of the Belgian Production Network*, International Productivity Monitor, Vol. 32, pp. 84-96

5.2. Weights

The data was aggregated for the same reference year as the firm characteristics dataset, ensuring temporal consistency between both sources. Although the original E-Fatura dataset is available at a monthly level, the relevant feature extracted corresponds to the *valor tributável agregado*—the total taxable value of transactions—summed over the entire year. This value is presented as a single column total per buyer-supplier pair.

Traditionally, datasets of this nature often rely on a binary variable to indicate whether a link exists between two entities (e.g., a value of 1 if a buyer is linked to a supplier, and 0 otherwise). However, by incorporating the aggregated transaction value, the dataset gains an enriched dimension. Instead of merely indicating the presence or absence of a relationship, the data now reflects the intensity or strength of the connection.

This enhancement adds considerable value to the machine learning pipeline. Not only can the algorithm classify whether a link exists, but it can also learn from the magnitude of past interactions—providing more nuanced signals during training. This continuous variable supports regression-based modelling, allowing for better generalization and more accurate predictions when inferring the likelihood or strength of a relationship in new, unseen data. The result is a more informative and versatile dataset, capable of supporting both classification and regression tasks.

It is important to be aware that the weights can have negative values and feature engineering this variable may be necessary. Negative values exist due to the yearly aggregated commercial balance between the pair of companies, for instance a series of payment adjustments from previous years. This negative links represents under 0,21% of total records in the dataset.

5.3. Firms' characteristics

To effectively model a buyer–supplier network, it is essential that each node — representing a firm — is characterised using variables that reflect its economic size, sector of activity, geographic location, and investment capacity.[8] Below, we highlight the importance of some of the key variables selected:

- NACE (NACE_SEC and NACE cod): These variables identify the sector of economic activity of each firm, which is crucial in understanding potential supply and demand relationships. For example, a manufacturing firm (industry) is more likely to be a supplier to a retail or services firm than the reverse. Sector classification also allows for stratified analyses and segmentation within the network.
- NUTS3: The regional location of each firm (at NUTS-3 level) enables the integration of spatial dimensions into the network. This is particularly relevant for identifying local supply chains or regional clustering effects, as well as for understanding transport and logistical constraints.
- TO (Turnover) and PURCH (Purchases): These variables provide insights into the economic volume of the firm. Turnover reflects the firm's sales capacity and market presence, while

[8] Here we follow approaches in the literature, e.g. Mungo L, Lafond F, Astudillo-Estévez P and Doyne Farmer J 2023 Reconstructing production networks using machine learning *J.Econ. Dyn. Control* **148** 104607 and Mungo, L., A. Brintrup, D. Garlaschelli and F. Lafond, Reconstructing supply networks ,*J. Phys. Complex.* **5** (2024) 012001.

Purchases may be a proxy for input dependency — i.e., how reliant the firm is on suppliers, which may indicate the potential number and strength of incoming links.

- NPE (Number of Persons Employed): Serves as a direct indicator of firm size and capacity, often associated with production volume, internal complexity, and likelihood of multiple business relationships.
- DIM (Company Dimension): This composite indicator, based on employment, turnover, and assets, allows for a standardised classification of firms (e.g., micro, small, medium, large), which can be useful for both network stratification and for identifying asymmetries in buyer–supplier relationships.
- FTA and FTI (Fixed Tangible Assets and Investments): These reflect the firm's capital intensity and long-term investment behaviour, which may be relevant to understand its role in the supply chain (e.g., asset-heavy manufacturing vs. light services).
- IAG and IAI (Intangible Assets and Investments): These offer insights into a firm's intellectual capital and innovation capacity, which may affect its position in the value chain, especially in sectors where know-how and R&D play a central role.

These variables, taken together, provide a multi-dimensional profile of each enterprise, allowing not only for link prediction but also for richer structural analysis of the buyer–supplier network.

The detailed variables are described in Table 1.

5.4. Dealing with spatial clustering

Spatial clustering and autocorrelation are common in urban network data, especially in densely populated areas. Although predictable in spatial interaction models (Griffith & Chun, 2015), they can lead to an over-biasing of short-distance connections. To counteract this, the following approaches are used (see Annex 2 for the references):

1. Log-Transformation of Distance Variables: Reduces the dominance of short distances and allows for better capture of long-distance interactions (Chun & Griffith, 2011).
2. Spatial Lag Models: They account for the influence of neighbouring units through a dependent variable with a spatial lag (Darmofal, 2015). Tan et al. (2025) emphasise the importance of properly parameterising the weight matrix.
3. Eigenvector Spatial Filtering (ESF): They deactivate autocorrelation by adding orthogonal eigenvectors as auxiliary variables. The effectiveness of ESF in flow data is confirmed by Chun and Griffith (2011), and Griffith and Chun (2015) point to its advantages over classical models.
4. Distance Decay Functions: Inverse distance and exponential decay allow for realistic modelling of spatial decay (Elhorst et al., 2024; Yuan et al., 2024). Tan et al. (2025) demonstrate that joint estimation of decay parameters increases the accuracy and robustness of models.
5. Generative modelling and inductive biases: As Chen (2025) points out, spatially aware architectures allow models to learn local and nonlocal dependencies, improving robustness against overfitting.
6. Modelling multiscale and heterogeneous networks: They integrate different types of relationships and spatial scales. Mann et al. (2023) and Yuan et al. (2024) demonstrate that such approaches improve connectivity prediction and account for autocorrelation in complex datasets.

7. Validation against known biases: Lamboley and Fourcade (2024) show that there is no universal filtering distance, emphasising the need for adaptive validation and sensitivity control strategies.

These approaches collectively increase the robustness of spatial models by enabling them to capture both local clusters and broader spatial dependencies.

6. Deriving training and test sets

The creation of training and test sets requires a structured, multi-stage workflow to ensure data quality, model reliability, and interpretability of results. This process involved several key steps, each contributing to the development of a robust and well-validated classification model. The stages included:

- **Formulation of the Model** – defining the modeling objective and specifying the dependent and explanatory variables;
- **Search, Classification, and Use of Features** – identifying relevant input variables and organizing them into meaningful feature groups;
- **Data Preprocessing** – validating, cleaning, and transforming the data to ensure consistency and suitability for modeling;
- **Model Development** – training multiple candidate models using the prepared dataset;
- **Model Selection – Validation by Metrics** – comparing model performance using standard classification metrics, with a focus on imbalanced data evaluation;
- **Feature Selection – Validation by SHAP** – interpreting model predictions and identifying key features using SHAP (SHapley Additive exPlanations) values.

For the current document, the first three stages are important, since they form the link between the available Portuguese data and the actual modelling and ML activities. Below, these are described in more detail, also showing that the link between the available data and the training software pipeline has been tested in practice.

These steps provided a systematic framework for constructing training and test sets that support reliable, explainable, and generalizable model outcomes.

FORMULATION OF THE MODEL

Based on the provided dataset, work was undertaken to identify explanatory and dependent variables and to formulate a model.

Assume we are investigating the impact of company characteristics on the transaction value Y_{ijt} between company *i* and company *j* at time *t*:

$$Y_{ijt} = \beta_1 X_{ijt} + \alpha_i + \alpha_j + \alpha_t + \varepsilon_{ijt}$$

where:

- X_{ijt} – explanatory variables, e.g., geographical distance, industry similarity, common suppliers,
- α_i, α_j – fixed effects for the sender and recipient companies,

- α_t – fixed effect for the given period,
- ε_{ijt} – random component.

We want to predict a binary variable (Y_{ijt}), which takes the value:

- 1 if there is a transactional/logistical connection between company *i* and company *j* at time *t*,
- 0 if there is no such connection.

This approach allows for the effective modelling of relationships while accounting for the unobserved heterogeneity across companies and time periods, which is essential when dealing with complex business networks.

SEARCH, CLASSIFICATION, AND USE OF FEATURES

Based on the provided dataset, variables can be classified based on whether they remain constant or change over time for a given company. The classification is as follows:

Variables with a Constant (or Almost Constant) Nature:

- ID – Company identifier (unchangeable).
- NACE_SEC – Sector of activity according to NACE (usually constant, changes rarely, e.g., due to restructuring).
- NUTS3 – Company location (change possible, but rare).
- DIM – Company dimension (based on long-term features, though it may change, but less frequently than annual financial indicators).

Variables with a Dynamic Nature (Changing Over Time):

- TO (Turnover) – Company revenue changes depending on the market situation.
- PURCH (Purchases) – Purchases of goods and services fluctuate.
- NPE (Number of Persons Employed) – The number of employees may change from year to year.
- ACT (Assets) – Company assets may change due to investments, depreciation, etc.
- WAGES (Expenditure on Personnel) – Expenditures on wages change over time (e.g., due to raises, changes in employment).

Those identified variables can be used in three ways:

1. Individual Company Features (constant or slow-changing variables, e.g., NACE sector, NUTS3).
2. Relationship features between Companies (e.g., size differences between companies, sector similarity).
3. Dynamic features (e.g., current turnover, number of employees).

This framework allows for an effective modelling approach, accounting for both stable company characteristics and dynamic, time-varying factors that influence transactional and logistical relationships between firms.

DATA PREPROCESING

The purpose is to prepare and integrate firm-level and transaction-level data in order to construct a dataset suitable for downstream analysis and machine learning.

Two data files were given. First one is a dataset containing buyer–supplier transaction records and the second - a dataset containing detailed firm-level information.

Two successive merge operations were conducted to enrich the transaction data (b2b_df) with firm-level attributes from business_df:

- **First Merge (Supplier Data)**:
  The transaction table is merged with firm data on the SUPPLIER identifier from b2b_df and the ID column from business_df.
  This join adds firm-level features for suppliers to each transaction record. The resulting suffix _sup is applied to these columns to indicate their origin.
- **Second Merge (Buyer Data)**:
  The enriched dataset is then merged again with business_df, this time using the BUYER column from b2b_df and ID from business_df.
  This step adds buyer-specific firm attributes to each record, using the suffix _buyer.

This two-step merging process produces a unified dataset containing transaction-level information with corresponding firm-level attributes for both parties involved in each relationship.

Following the merge operations, the temporary columns ID_sup and ID_buyer, which were created to support the join logic, are removed from the dataset to streamline the table.

Additional Preprocessing Steps Performed:

- Validation of data integrity following dataset merging
- Verification of data types across all columns
- Identification and removal of duplicate records
- Detection of negative values in numeric fields
- Detection of missing values
- Imputation of missing values with zeros (restricted to the *LINKED* column)
- Identification of zero values
- Verification that values in the *NACE sector* column contain exactly one alphabetical character

Recommendations for Further Work:

- Normalize the values in the *DIST* column using min-max scaling and visualize the resulting distribution
- Perform undersampling in the *LINKED* column to address class imbalance

During actual training of a specific model, the selected and processed datasets in use will be stored. This enables finetuning of the model with e.g. investigating the optimal model parameter settings and testing the potential sensitivity of the model for minor changes in the dataset. When datasets for additional years become available, the sensitivity can also be tested by running the same models on an earlier or later dataset.

7. Validation of the reconstructed networks

Validation of the modelling results are part of the next phase of the work, but we can already mention some of the validation steps which we intend to use. For validating the unweighted reconstructed network, an overall plausibility check will be done by calculating a number of standard descriptives for this type of networks and compare the results with other networks.[9] Validating the weighted networks will be done with this procedure: aggregate the firm-level dataset to the industry x industry input output table and calculate the ratios in each cell between the reconstructed network

[9] Talaga, S. and A. Nowak, *Structural measures of similarity and complementarity in complex networks*, Scientific Reports | (2022) 12:16580 and Bacilieri, A., Borsos, A., Astudillo-Estevez, P., Hoefer, M., & Lafond, F. *Firm-level production networks: What do we (really) know?,* INET Oxford Working Paper No. 2023-08.

and the IO table. Due to conceptual differences the ratios are not supposed to be one but they should be in a certain value range to be plausible. These value ranges per cell will be derived from the ratios of the Portuguese network dataset and IO table.

8. Summary

The document outlines the creation of training and test datasets for a machine learning model aimed at reconstructing buyer-supplier networks among enterprises. The project, part of the AIML4OS initiative, involves integrating administrative firm data and transactional links from electronic tax invoices to form a comprehensive dataset. This integration allows for detailed economic and policy analyses by leveraging machine learning techniques.

The report highlights the methodology used to create the firm-level network dataset, starting with the identification of relevant data sources in the Portuguese Statistical Office. Two key sources were selected: detailed enterprise characteristics (SCIE) and tax invoice records (e-Fatura). The enterprise data includes information such as firm size, sector of activity, and geographical location, while the tax invoice data provides direct links between firms, identifying supplier-buyer relationships. The data was cleaned and pre-processed to ensure accuracy, and the resulting dataset was enriched with additional firm characteristics like the physical distance between buyers and suppliers.

A synthetic dataset was also created to enable collaboration without compromising confidentiality. This dataset was generated using the {synthpop} R-package and includes a subset of economically significant firms. The synthetic data preserves the relationships between variables while protecting the original data's confidentiality. The report emphasizes the importance of understanding the limitations of the synthetic dataset, particularly its bias towards larger firms and regional imbalances.

Finally, the document discusses the challenges and methodologies involved in calculating link characteristics, such as geographic distances between firms. It also addresses the importance of sectoral and regional representation in the dataset to ensure the model's generalizability. The integration of these diverse data sources and careful preprocessing steps aim to create a robust foundation for developing predictive models that can accurately reconstruct buyer-supplier networks.

Supporting information

All the graphs and supporting images in this report are part of the technical documentation available on the Onyxia platform as detailed below.

The synthetic firm characteristics data is available to project members in Onyxia platform:
https://minio.lab.sspcloud.fr/projet-aiml4os-wp11/aiml4os_wp11_synthetic_data_1/wp11_companies_synthetic_data.parquet

The synthetic b2b data is available to project members in Onyxia platform:
https://minio.lab.sspcloud.fr/projet-aiml4os-wp11/aiml4os_wp11_synthetic_data_1/wp11_b2b_synthetic_data.parquet

The analysis by sections is available to project members in Onyxia platform:
https://minio.lab.sspcloud.fr/projet-aiml4os-wp11/NACE sector analysis/NACE_distribution.html

The synthetic dataset characteristics is available to project members in Onyxia platform:
https://minio.lab.sspcloud.fr/projet-aiml4os-wp11/Synthetic_Dataset_Characteristics/fc_pt.html

Annex 1 - Variables of Portuguese electronic invoices

E-Invoices (V_TF_EFAT_ENCRIPTADA_AAAA)

Atributo	Tipo	Descritivo	EN
ANO	TEXTO (4)	Ano de emissão de fatura	Year of invoice issuance
MES	TEXTO (2)	Mês de emissão de fatura	Month of invoice issuance
VERSAO	NUMERO	Versão dos dados (relativo ao ano, mês)	Data version (relative to year, month)
ID_SEQ	NUMERO	Número sequencial de registo (relativo ao ano, mês)	Sequential registration number (relative to year, month)
NIF_EMITENTE	TEXTO (9)	Número de Identificação Fiscal da entidade (singular ou coletiva) que emitiu fatura	Tax identification number of the entity (individual or collective) that issued the invoice
NIF_ADQUIRENTE_NAC_COL	TEXTO (9)	Número de Identificação Fiscal da entidade coletiva (e nacional) adquirente	Tax Identification Number of the acquiring collective (and national) entity
NIF_ADQUIRENTE_NAC_SING_ENCR	TEXTO (64)	Número de Identificação Fiscal encriptado da entidade singular adquirente; Inclui também os NIF 999999990 que representam faturação nacional com ausência de NIF, por forma a manter num mesmo atributo os NIF adquirentes singulares	Encrypted Tax Identification Number of the acquiring individual entity; It also includes NIF 999999990 that represent national invoicing without a NIF, in order to maintain the NIF of individual acquirers in the same attribute.
NIF_ADQUIRENTE_ESTR	TEXTO (200)	Número de Identificação Fiscal de entidades adquirentes estrangeiras	Tax Identification Number of foreign acquiring entities
VALOR_TRIBUTAVEL	NUMERO	Valor tributável (correspondente ao valor do adquirente agregado no mês, para um determinado emitente)	Taxable value (corresponding to the aggregate acquirer value in the month, for a given issuer)
TIPO_VALOR_TRIBUTAVEL	TEXTO (1)	Identifica o tipo de valor tributável (por defeito='O', de Original); Descodifica com TD_TIPO_VALOR_TRIBUTAVEL	Identifies the type of taxable value (default='O', for Original); Decode with TD_TIPO_VALOR_TRIBUTAVEL
TIPO_EMITENTE	NUMERO	Identifica se a entidade emitente é do tipo Singular ou Coletivo; Descodifica com tabela TD_TIPO_EMITENTE	Identifies whether the issuing entity is of the Individual or Collective type; Decode with table TD_TIPO_EMITENTE
TIPO_MERCADO	NUMERO	Identifica o tipo de mercado que adquiriu; Descodifica com TD_TIPO_MERCADO	Identify the type of market you acquired; Decode with TD_TIPO_MERCADO
PAIS_DSG_SMI	TEXTO (200)	Designação do país adquirente	Designation of the acquiring country

PAIS_COD_SMI	TEXTO (5)	Código ISO Alpha 2 do país adquirente	ISO Alpha 2 code of the acquiring country
NUTIII_EMITENTE	TEXTO (3)	NUTSIII do Emitente	NUTSIII of the Issuer
STA	TEXTO (2)	Situação perante a atividade do Emitente	Situation regarding the Issuer's activity
CAE3	TEXTO (5)	Código de atividade económica (CAE Rev3) do Emitente	Economic activity code (CAE Rev3) of the Issuer
FJR	TEXTO (3)	Código da forma jurídica do Emitente	Code of the Issuer's legal form
SIN	TEXTO (10)	Código do Setor Institucional (SIN) do Emitente	Institutional Sector Code (SIN) of the Issuer
ZONA_FRANCA	TEXTO (1)	Código de Zona Franca do Emitente	Issuer Free Zone Code
DDCCFF_EMITENTE	TEXTO (6)	Código DDCCFF do Emitente	Issuer Code DDCCFF (concatenation of district, municipality and parish)
FONTE_CARACTERIZACAO_EMITENTE	TEXTO (2)	Identifica a fonte de dados usada para caracterização do Emitente; Descodifica com TD_FONTE_CARACTERIZACAO	Identifies the data source used to characterize the Issuer; Decode with TD_FONTE_CARACTERIZACAO
CLASSE_ADQUIRENTE	TEXTO (2)	Tipifica o Adquirente, com base no seu NIF e origem; Descodifica com TD_CLASSE_ADQUIRENTE	Types the Purchaser, based on their NIF and origin; Decode with TD_CLASSE_ACQUIRENTE
DTCCFF_ADQUIRENTE	TEXTO (6)	Código DDCCFF do Adquirente	Purchaser Code DDCCFF (concatenation of district, municipality and parish)
FONTE_CARACTERIZACAO_ADQUIRENTE	TEXTO (2)	Identifica a fonte de dados usada para caracterização do Adquirente; Descodifica com TD_FONTE_CARACTERIZACAO	Identifies the data source used to characterize the Acquirer; Decode with TD_FONTE_CARACTERIZACAO
NUTIII_2024_EMITENTE	TEXTO (3)	NUTSIII (versão 2024) do Emitente	NUTSIII (version 2024) of the Issuer

Annex 2 References spatial clustering

1. Chen, Z. (2025). Rethinking inductive bias and generative modelling losses for geographically neural network weighted regression. arXiv preprint. https://arxiv.org/abs/2507.09958v2 [cs.LG]

2. Chun, Y., & Griffith, D. A. (2011). Modeling network autocorrelation in space-time migration flow data. Annals of the Association of American Geographers, 101(3), 523–536. https://www.jstor.org/stable/27980199

3. Darmofal, D. (2015). Spatial analysis for the social sciences. Cambridge University Press. https://doi.org/10.1017/CBO9781139051293

4. Elhorst, J. P., Tziolas, I., Tan, C., & Milionis, P. (2024). The distance decay effect and spatial reach of spillovers. Journal of Geographical Systems. https://link.springer.com/article/10.1007/s10109-024-00440-5

5. Griffith, D. A., & Chun, Y. (2015). Spatial autocorrelation in spatial interactions models. Networks and Spatial Economics, 15(3), 711–730. https://link.springer.com/article/10.1007/s11067-014-9256-4

6. Lamboley, Q., & Fourcade, Y. (2024). No optimal spatial filtering distance for mitigating sampling bias. Journal of Biogeography. https://onlinelibrary.wiley.com/doi/10.1111/jbi.14854

7. Mann, G., Dsouza, A., Yu, R., & Demidova, E. (2023). Spatial link prediction with spatial and semantic embeddings. In International Semantic Web Conference (ISWC). https://link.springer.com/chapter/10.1007/978-3-031-47240-4_10

8. Tan, C., Kesina, M., & Elhorst, J. P. (2025). Parameterizing spatial weight matrices in spatial econometric models. Political Analysis, 33(1), 49–63. https://doi.org/10.1017/pan.2024.16

9. Yuan, J., Zhao, Y., Yi, D., Jin, S., Zhou, H., & Zhang, J. (2024). Exploring multi-relational spatial interaction imputation with distance-decay effects. International Journal of Digital Earth. https://www.tandfonline.com/doi/full/10.1080/17538947.2023.2300316