Analysis

Mapping Physical Climate Risk Exposure in Euro Area firms: Linking E-PRTR and RIAD with a Fuzzy Match

Flavio De Carolis, Kitty Rang & Michiel Nijhuis

DeNederlandscheBank
EUROSYSTEEM

Mapping Physical Climate Risk Exposure in Euro Area firms: Linking E-PRTR and RIAD with a Fuzzy Match

©2025 De Nederlandsche Bank NV

Authors: Flavio De Carolis, Kitty Rang and Michiel Nijhuis. Special thanks to Justin Dijk, Patty Duijm, Iman van Lelyveld, Julika Herzberg, Malgorzata Osiewicz and Francesca Rinaldi for their valuable feedback. All remaining errors are ours.

With the 'DNB Analysis' series, De Nederlandsche Bank aims to provide insight into the analyses it conducts for current policy issues. The views expressed are those of the authors, and do not necessarily reflect the official views of De Nederlandsche Bank. No part of this publication may be reproduced and/or published by means of print, photocopy, microfilm or by any other means, nor may it be stored in a retrieval system, without the prior written permission of De Nederlandsche Bank.

De Nederlandsche Bank NV
P.O. Box 98
1000 AB Amsterdam
Internet: www.dnb.nl
Email: info@dnb.nl

Abstract

We develop a methodology to assess the exposure of euro area companies to physical climate risks by incorporating the geographic distribution of their production facilities. We apply this approach by linking company-level data from the Register of Institutions and Affiliates Data (RIAD) with facility-level data from the European Pollutant Release and Transfer Register (E-PRTR) using a fuzzy string matching algorithm enhanced with machine learning. The methodology is applied to estimate expected annual losses (EAL) from floods and windstorms, following the framework developed by the European System of Central Banks. Our findings show that accounting for local units significantly affects the quantification of flood risk exposure, for both river and coastal floods, while the impact on windstorm risk is limited. We further analyze the spatial distribution of EAL differences across NUTS 3 regions, highlighting areas where relying solely on headquarter locations leads to substantial underestimation of risk.

1 Introduction

The economic losses triggered by extreme weather events in Europe have increased in recent decades. The European Environment Agency reports that in the last four decades (1980-2022) losses related to weather and climate-related extremes in Europe totaled 650 billion, of which 53.2 occurred in 2022 alone (EEA, 2023). Although the losses vary strongly from one year to another, the 30-year moving average indicates an upward trend. Furthermore, the exposure of losses is heterogeneously distributed throughout Europe, not only depending on the type of extreme weather event, but also on the insurance rate as insurance coverage ratios differ from country to country (EIOPA, 2022).

The exposure of the financial system to physical climate risks is of great relevance for the European System of Central Banks (ESCB) and to other market participants due to the potential impact that extreme weather events have on asset prices, financial stability and the implementation of monetary policy. For instance, several papers have investigated the effect of climate events on asset prices. Braun et al. (2024) find evidence for a large and statistically significant hurricane premiums for firms that were affected by a hurricane. On the other hand, Hong et al. (2019) find that stock markets do not sufficiently incorporate information about drought trends in asset prices. Additionally, the ESCB Statistical Committee releases statistical climate risk indicators and the respective methodologies on an annual basis to improve awareness and facilitate the analysis of exposure to physical climate risk within the financial system (ECB, 2023, 2024). The climate risk indicators aim to capture potential losses for the financial system by quantifying the expected annual losses for the non-financial companies they invest in. The methodologies to estimate the exposure of companies to floods and windstorms derived by the ESCB are based on the location of assets together with regionally calibrated damage functions to estimate the expected annual losses on a company level. However, the location used in the analysis is based mainly on the exposure of companies' headquarters to floods and windstorms. The location of these headquarters, however, might not coincide with the location of most of the company's assets.

The current finance literature shows significant improvements in the precision of loss estimates by including information on production facilities in financial analysis. For example, Bressan et al (2024) show that climate-related losses faced by European Investors in publicly traded equities of Mexican issuers are underestimated by 70% when ignoring asset-level information. Atlan et al. (2024) build on the work of the ESCB Statistical Committee by analyzing the impact of including local units of French companies on the flood risk and drought indicators. The results do not indicate a structural underestimation of the physical risk at the aggregate level. However, for individual companies, the risk profile of the headquarter location and its local units can differ significantly. On the other hand, Loberto & Russo (2024) find that around 17% of the companies that were impacted by the flood of May 16-17 2023 in Emilia-Romagna have a headquarter outside of the region.

Multiple studies have worked on linking the European Pollutant Release and Transfer Register (E-PRTR) to company registers. While Germeshausen et al. (2022) linked the E-PRTR data to financial information using Orbis from Bureau van Dijk, we extend their methodology by matching the data set with the Register of Institutions and Affiliates Data (RIAD) and using a machine learning algorithm to improve the fuzzy string matching. Additionally, Bauer et al. (2025) merge publicly available local unit data with companies' headquarters information to analyze how local institutional ownership shapes price discovery around extreme weather events. We contribute to the literature in two ways: first, we provide the first Europe wide analysis investigating the impact of local units on the assessment of physical risk exposure using the ESCB methodology (Atlan et al., 2024; Loberto & Russo, 2024), second we replicate the name matching algorithm developed in Bauer et al. (2025) using central bank data on companies and providing free access to the name matching algorithm.

1.1 Downside of current approach and scope of the analysis

The methodology developed in ECB (2023, 2024) uses the addresses provided in the Register of Institutions and Affiliates Data (RIAD) as the location of exposure of companies to physical risks. However, RIAD collects information at the level of the legal entity and of a single branch per country. Consequently, in the case of multiple locations of a single company (e.g. production sites that are at a different location from the headquarters), the information necessary for a full assessment of physical risk is not available in this dataset.

We overcome this issue by exploiting publicly available production facility data, such as the European Pollutant Release and Transfer Register (E-PRTR) provided by the European Commission. Due to the absence of identifiers to link the E-PRTR and RIAD data, we enhance the company string matching algorithm by Nijhuis (2022). We then replicate the methodology of ECB (2023, 2024) for floods and storms, including the location of linked production facilities, and investigate the difference in the indicators by NUTS 3 regions. We perform both the name matching and the calculation of losses on the single entity level, meaning that we do not take into account group structure.

We find that adding information on local units (LU) is relevant for river and coastal floods but less for windstorms: the average Expected Anual Loss (EAL) for local units due to floods can be up to 10 percentage points higher than that of the headquarter. Furthermore, while for windstorms the results are more aligned in Europe, some regions are more exposed than others to river and coastal floods, showing that for some regions it is more important than others to account for LU. For instance, companies with headquarters in urban centers such as Madrid and Paris experience on average, respectively, up to -6.28 and -3.5 percentage points difference in EAL for river floods when accounting for local units. The setup of the paper is as follows: in Section 2, we explain the fuzzy string matching algorithm and the ESCB methodology for physical risk, in Section 3 we present the results of the analysis and in Section 4 we explain limitations and extensions of our work, thus concluding the analysis.

2 Methodology

2.1 The name matching algorithm

Merging large data sets when facing inconsistencies in company names requires an approach that balances efficiency and accuracy. In an attempt to extend the methodology proposed by Nijhuis (2022), we implement several machine learning algorithms that take advantage of different measures of similarity to evaluate the match between the company name in the two data sets.

While preserving the use of cosine similarity in combination with various more complex similarity measures as outlined in Nijhuis (2022), we also enhance the approach by incorporating a machine learning model that optimizes the weighting factors of the complex similarity measures for improved matching accuracy. The enhanced approach consists of four steps: the initial data cleaning, the identification of potential matches using cosine similarity matching on the whole dataset, the calculation of similarity scores using different similarity measures, and a machine learning classification of the possible matches. Figure 2.1 gives a schematic overview of these steps. In the following, the four steps are explained in more detail.

2.1.1 Step 1: Data cleaning

The data cleaning consists of the transliteration of non-ASCII characters, the removal of punctuation, the removal of legal prefixes and suffixes and the removal or abbreviation of common terms like "International". More information on the pre-processing stage can be found in Appendix B.

2.1.2 Step 2: Cosine similarity matching

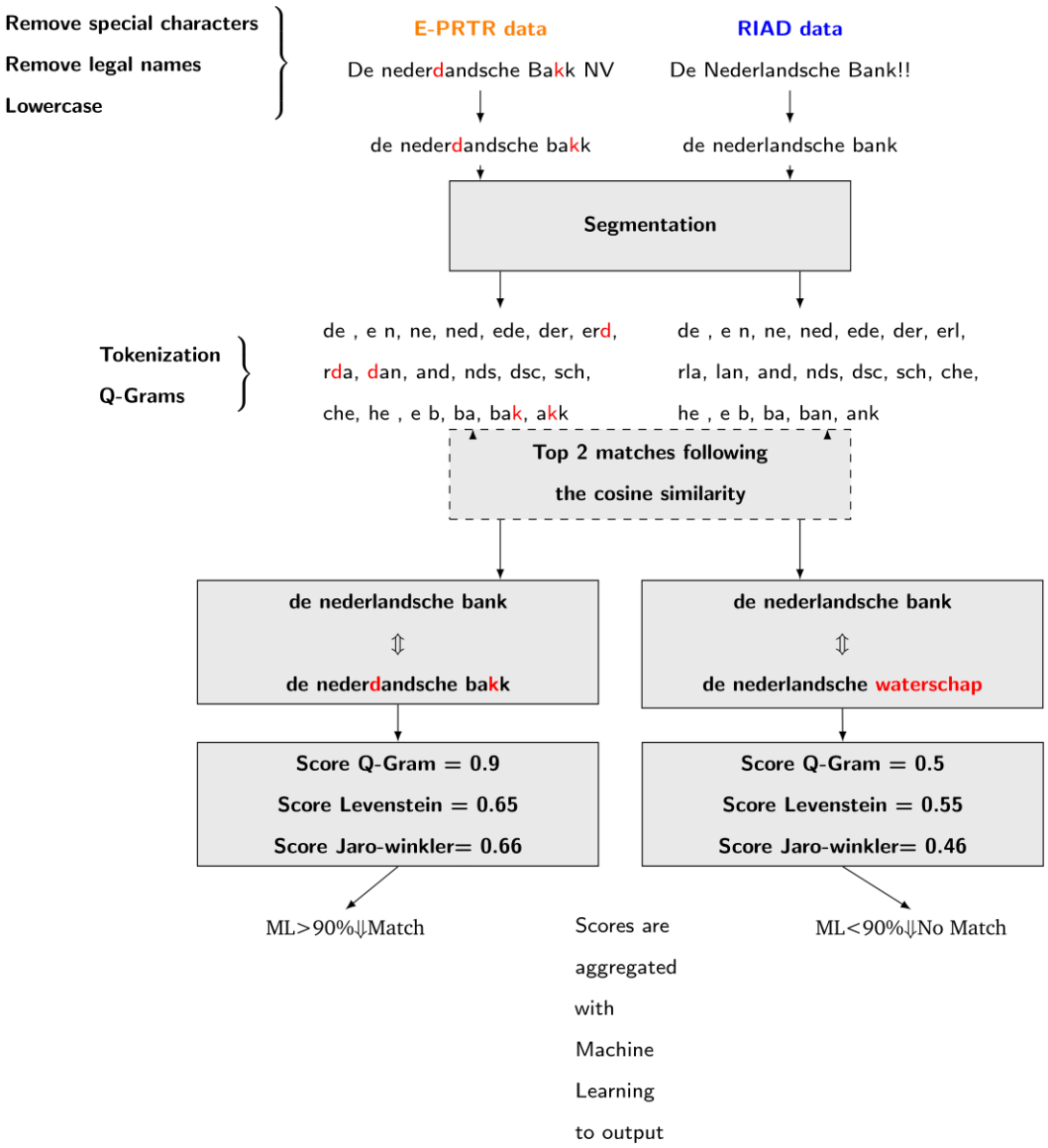
The second step involves quickly identifying potential matches within the entire dataset to limit processing times in the next steps of the name matching. The identification of potential matches begins by generating trigrams (all chunks of three consecutive characters) for each name in both datasets (see Figure 2.1). Next, a term frequency-inverse document frequency (TF-IDF) algorithm is applied to the trigrams, which reduces the weight of commonly occurring ones. Afterward, the cosine similarity is calculated between a name in one data set and all names in the other dataset (see Figure 2.1). For each name in E-PRTR we select the 50 most similar names in RIAD in terms of cosine similarity. The number of potential matches is limited to 50, as the similarity score of the last match is typically lower than 0.3, indicating minimal similarity.

2.1.3 Step 3: Calculation of similarity measures

Although cosine similarity is commonly used for name matching, it is a limited way to determine whether two strings match. Many kinds of variation are not distinguished by the cosine similarity on the trigrams. Take for instance the names $sara$ and $sarah$. The trigram decomposition of the names are $[sar, ara]$ and $[\mathrm{sar,ara,rah}]$. The only overlapping trigram is ara, leading to a relatively low cosine similarity. However, the addition of one letter transforms $sara$ into $sarah$, meaning that the edit-based distance is small.

To capture different ways in which strings can be similar, we calculate three different similarity metrics for each of the 50 potential matches, namely the Levenshtein distance, the Jaro-Winkler similarity and the Q-gram distance. Each of these three similarity measures addresses different aspects of string comparison (Van Der Loo, 2014). More information on the calculation of the similarity measures can be found in Appendix B.

Figure 2.1 The Name Matching algorithm in a chart: A top-bottom visualization of the algorithm.



2.1.4 Step 4: Machine learning

By applying the three algorithms to all potential matches, we obtain three scores per potential match. It still needs to be determined whether any of the potential matches is an actual match. To facilitate this determination, a machine learning approach is applied (see Figure 2.1). The machine learning approach takes the scores from the different similarity measures of a potential match as input and tries to classify whether a match is a true match or not. For this approach to work, there needs to be a training set of verified matches. A subset of 5% of

the data is manually matched and used to train the model and validate the results. This training dataset is highly imbalanced, as there is at most one true match amongst the 50 potential candidates. To address this imbalance, all actual matches are retained in the training data, while a strategically selected subset of non-matches—those with the highest similarity scores—is included. This approach, known as undersampling, helps create a more balanced training set, which is crucial for enabling the model to effectively learn to identify true matches. The resulting dataset is used to train a variety of machine learning models. Model performance is evaluated using cross-validation on the training set, as well as on an original, unbalanced test set. The use of the unbalanced test set provides a more realistic assessment of the model's performance, while the use of the balanced training set ensures that the recall of the trained model will be high enough. More information on the model selection and performance can be found in Appendix B.

2.2 The ESCB methodology for physical risk scores

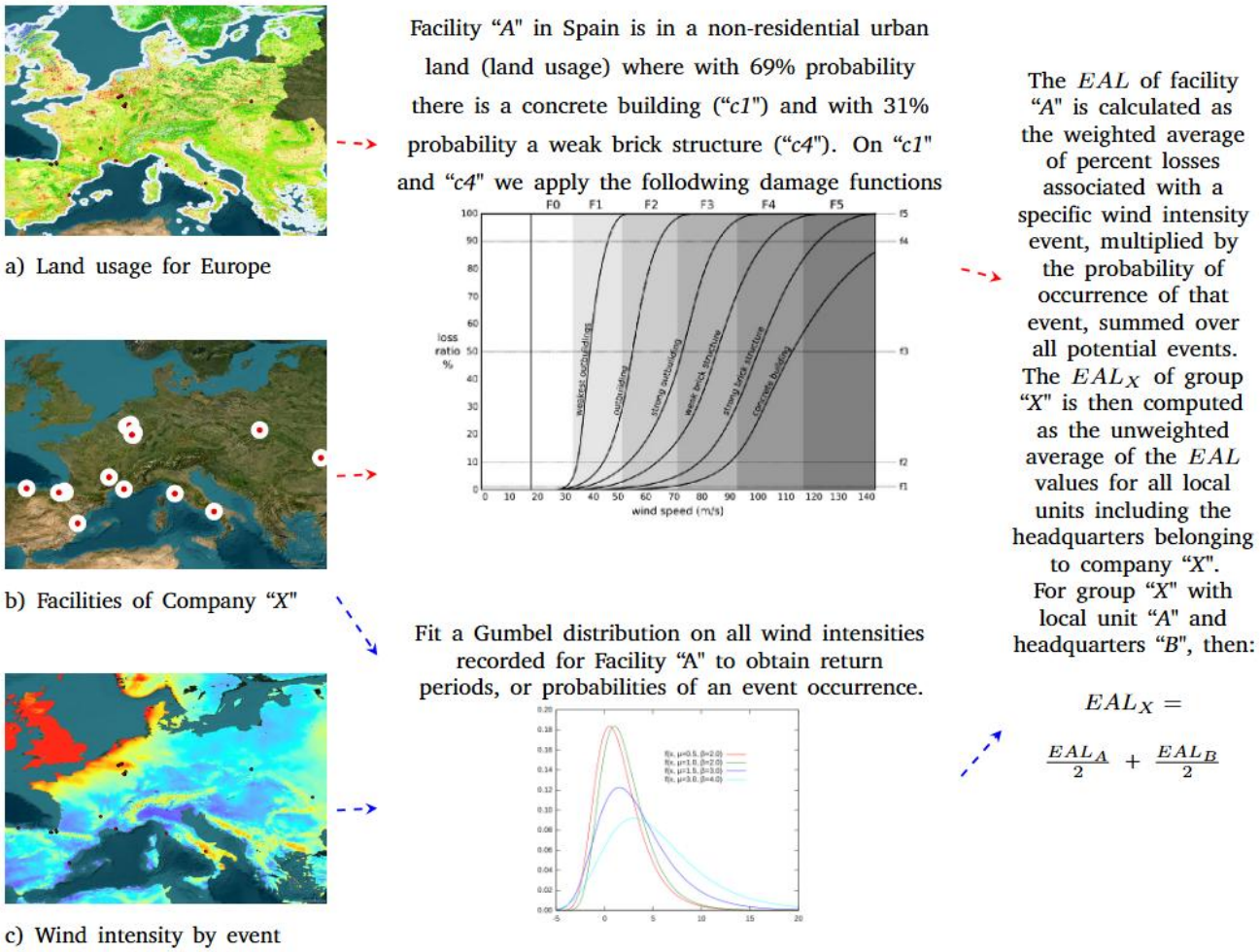
We replicate the ECB (2024) framework to calculate expected annual losses (EAL) using data from local units (production facilities), land use, extreme weather events, distribution maps of type of buildings and damage functions, as shown in Figure 2.2. LU data is sourced from the E-PRTR (mid-left panel of Figure 2.2), while the land use maps are provided by Copernicus Land Monitoring (upper left panel of Figure 2.2). Land use refers to the management and designation of land for specific activities, such as residential, commercial, agricultural, industrial, and recreational purposes.

To compute EAL from extreme weather events at LU level, we follow three main steps. First, we calculate the probabilities of events with specific intensities occurring at the local unit's location. Second, we identify the damage functions associated with these intensities to estimate the losses incurred by the facility. Third, we compute EAL as the weighted average of probabilities and losses for all potential events that could affect LU. Figure 2.2 is a reference to guide the understanding.

To compute EAL from storms, we first calculate the probability of an event with a given intensity occurring at the facility location within the next year. To achieve this, we use maps of the intensities of extreme weather events based on historical occurrences of storms (lower left panel of Figure 2.2). Specifically, we use historical footprint maps and for each pixel, we derive a distribution of wind intensity recordings. Return periods or probabilities of a storm occurring within a given wind intensity range in a year are calculated by fitting the historical distribution of events at each geographic pixel to a Gumbel distribution (lower middle panel of Figure 2.2)[1]. We then identify the level of damage that a local unit would incur from an event with a specific wind intensity. To do this identification, we require the type of building of the facility, the intensity of the wind event, and the associated loss. Since the E-PRTR does not provide information on the type of building of LU, we use a proxy suggested by ECB (2023, 2024) the combination of land use and the distribution of buildings by land use for each country in Europe. Furthermore, the damage functions for Europe are already calibrated for different types of buildings ECB (2023, 2024). An example of this step is provided in the upper middle panel of Figure 2.2. Consequently, the damage functions for storms at the local unit level result from the weighted average of the damage functions associated with the different types of buildings in the land use code of the facility location where the weights are given by the frequencies of different building types in that area. Given the probabilities of the occurrence of different events, the wind intensities of these events, and the damage functions just introduced, we can compute the EAL at the facility level as in ECB (2023, 2024) (right panel of Figure 2.2).

[1] The Gumbel distribution is commonly used to model the distribution of the maximum (or minimum) of a number of samples from various distributions. For example, in this context, it represents the distribution of maximum wind intensity in a specific region.

Figure 2.2 The relevant steps to compute the expected annual loss EAL from storms at a group level using the ESCB method[2]



To compute EAL from floods, we follow the same steps as for storms, albeit more straightforwardly. First, flood hazard maps provide event intensities linked to specific return periods and their associated probabilities of occurrence. These maps were sourced from Delft University (Paprotny et al., 2017). Second, flood damage functions are calibrated at the level of country and land use and are provided for different levels of flood depths (Huizinga et al., 2017). Third, using the location LU, which is associated with a specific land use (see the left middle and upper panels of Figure 2.2), we combine flood depth maps by return period with the respective damage functions to calculate the EAL for floods at the facility level.

[2] From left to right, we outline the steps to compute the EAL for group X based on the framework developed by the European System of Central Banks (ESCB) (ECB, 2023, 2024). Red dashed lines indicate the data sources used to compute damage functions for a specific event occurrence. Blue dashed lines represent the sources required to compute the probability of an event with a specific wind intensity, modeled using a Gumbel distribution, where the probability mass is concentrated on less extreme events. Finally, this information is combined to calculate the EAL at the facility level that is then aggregated to derive the company-level EAL.

3 Results

In this analysis, we compare the EAL obtained at the headquarter (HQ) level from single entities in the Euro Area of RIAD with the average EAL obtained using only local units (LU) derived from the E-PRTR database.[3] We analyze results for the Euro area distribution of EAL differences between HQ and LU at a company level. In a second step, we plot the differences using the NUTS 3 region of the companies and show for which regions and physical climate risks there are bigger differences in the Euro Area.

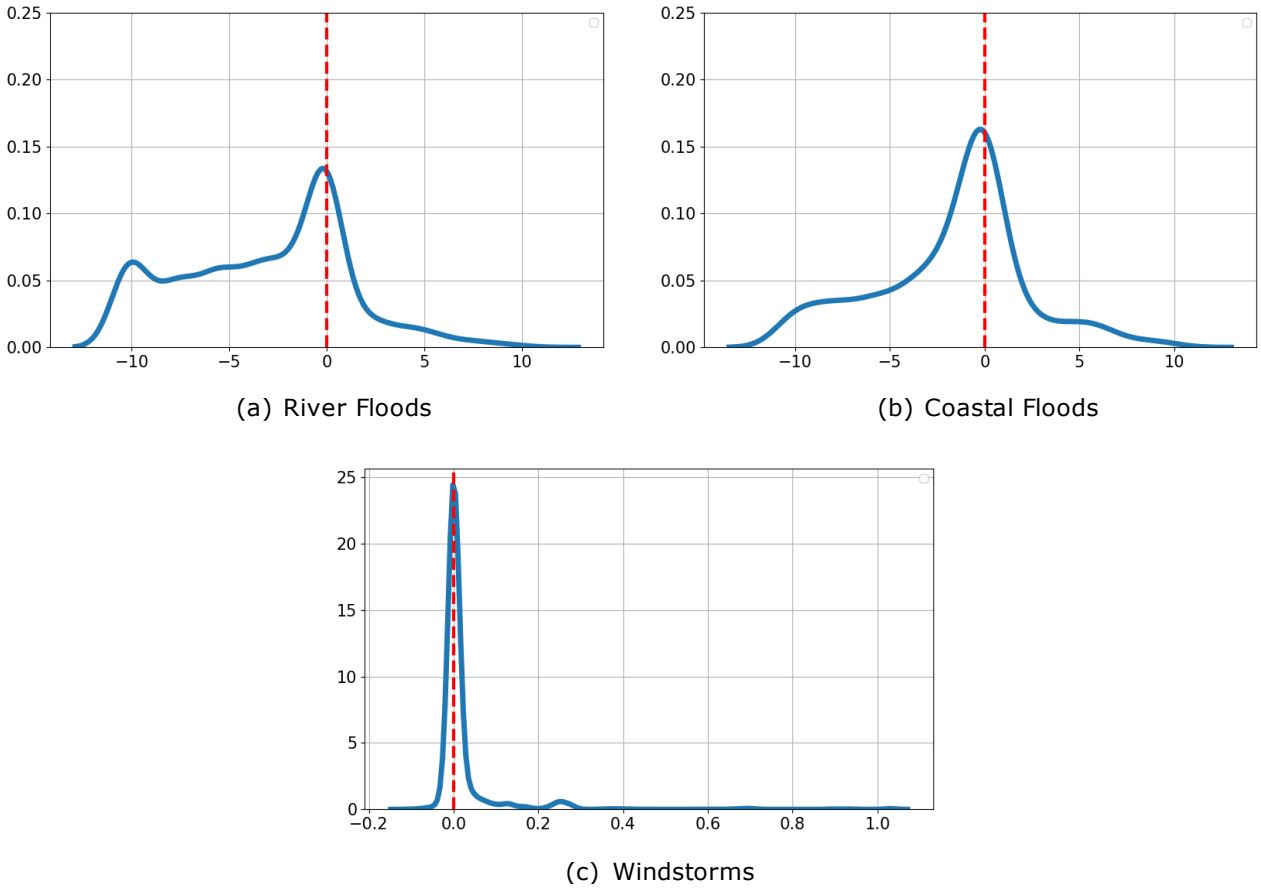
We link 43% of the E-PRTR facilities to RIAD single entities, applying expert judgment on potential matches to avoid the presence of false positives. We achieve this result starting with a sample of 11,600,142 RIAD entities and 63,106 E-PRTR facilities. The resulting sample is 27,022 E-PRTR facilities linked to 19,049 RIAD single entities.

Although we match 27,022 E-PRTR facilities to RIAD single entities, only a fraction of these facilities has an exposure to windstorms, river and coastal floods. Wind impacts broad regions; therefore, we compute an EAL from windstorms for 26,530 facilities. Floods impact regions near rivers or the coast, as such only 3,468 facilities are exposed to river floods, and 936 facilities to coastal floods. In the complete E-PRTR sample, we computed EAL for 97% of the sample for wind storms, 13% for river floods, and 4% for coastal floods. In the sample matched with RIAD the percents are 98%, 13% and 3%, respectively, thus mirroring the data for the full sample.

We identify significant differences for river and coastal floods and negligible ones for windstorms. In Figure 3.1 we plot the differences between HQ and LU for all matched LU. Negative values indicate that the EAL of the headquarter location is lower than the average EAL for all local units. For instance, a value of -1 means that the average EAL of the HQ is 1 percentage point lower than the average EAL of the LU. In this case, only including the headquarter location would lead to an underestimation of risk. For river floods, it is visible from Subfigure (a) of Figure 3.1 that LU has up to 10% percentage points higher EAL for LU compared to HQ. Similar findings are visible for coastal floods in Subfigure (b) of Figure 3.1. Finally, as far as windstorms are concerned, the EAL differences are largely distributed around zero or above, indicating that EAL computed at an HQ level are mostly higher than those computed at LU level.

[3] More information on RIAD, E-PRTR and the exposure risk maps is available in Appendix A.

Figure 3.1 The distribution of the EAL differences between HQ and LU [4]



(a) River Floods

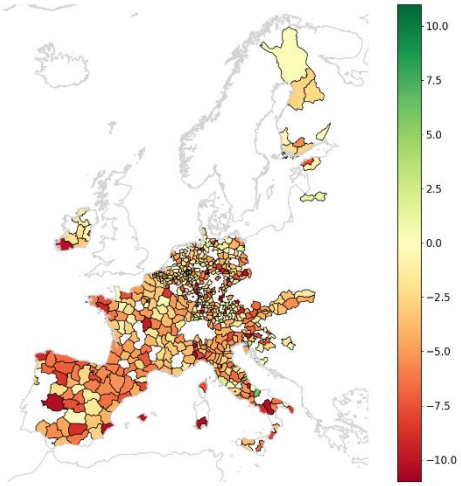
(b) Coastal Floods

(c) Windstorms

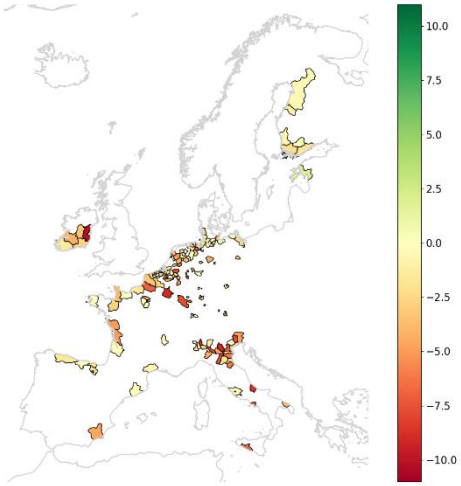
Zooming into NUTS 3 regions gives a geographical indication of which areas are more prone to differences between the EAL of the HQ and the EAL of the LU. In Figure 3.2 we present the map of NUTS regions. Negative values (red) indicate that the HQ located in that NUTS region have on average lower EAL than the LU associated with those HQ. For river floods, there are significant differences in some regions in France, Germany, Ireland, Italy, and Spain. For coastal floods, differences are mainly in located in Belgium, France, Germany, Ireland, Italy, and the Netherlands. We do not find differences for any specific region in windstorms.

[4] The density plot shows the difference between EAL at the HQ and the average at the LU level in percent (y-axis). 0.25 stand for % of observations. The frequency is based on the absolute number of RIAD companies linked to LU. Negative values show higher EAL estimates at the LU level.

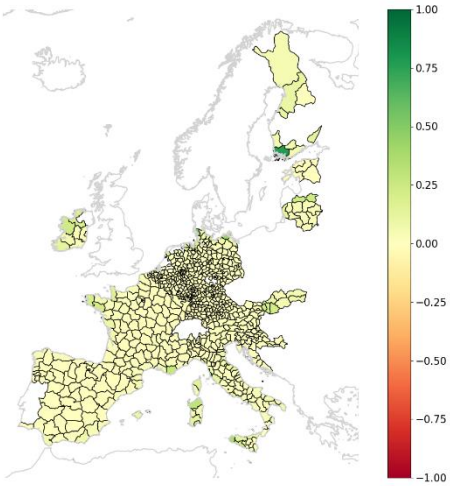
Figure 3.2 Euro area maps on NUTS 3 region level of differences in EAL estimates between HQ and LU [5]



(a) River Floods



(b) Coastal Floods



(c) Windstorms

[5] The map shows the difference in the estimates at a NUTS region level between HQ and the average between the LU. Red indicate higher EAL values for LU with respect to the HQ. The NUTS regions plotted is the one from the HQ and differences are in percentage points.

4 Discussion and conclusion

In this paper, we leverage the methodology developed by Nijhuis (2022) to investigate the advantages of accounting for local units (LU) in the calculation of expected annual loss (EAL) for different physical climate risks. To do so, we perform a fuzzy string matching between the companies available in RIAD and the production facilities in the E-PRTR. We then replicate the ESCB methodology to compute the exposure of companies to physical risk and find that the differences in the estimates of those companies that are linked to LU depend on the type of physical risk.

Storms and floods are different in their intrinsic risk nature. For example, storms affect wide regions, while river and coastal floods are more location-specific. Therefore, for storms, it is less relevant to know the exact location of single entities' production facilities as long as the headquarters are in the same country. In contrast, for river and coastal floods, location is crucial to know the exact exposure of the building. For instance, single entities with headquarters in urban centers such as Madrid and Paris experience, respectively, up to -6.28 and -3.5 percentage points difference for river floods when accounting for local units. Similarly, single entities located in the region of Mantova and Milan in Italy experience a difference of up to -4.5 and -2.9 percentage points for coastal floods when accounting for local units. Financial institutions investing in these regions would benefit from access to more granular information on the location of production facilities in the assessment of flood risk. Therefore, we advise the inclusion of local units in the analysis of climate risk for the financial sector, such as the climate-change related indicators of the ESCB (ECB, 2024).

In this paper, we provided an analysis based on ESCB data sources and methodology and found that LU provides useful information for the analysis on climate risk. In particular for flood risk, including additional information on local units leads to a better risk assessment. This analysis showcases one way of using the fuzzy string matching algorithm to analyze physical climate risk. Specifically, we use RIAD to identify companies and the E-PRTR to link facilities to it. The algorithm can be used more generally to link financial data sources with non-financial data sources to facilitate more in-depth analysis of physical climate risk. The Spatial Finance Initiative in Oxford provided an extensive review of countries that provide databases on production facilities, their location, as well as other information (Christiaen et al., 2025). Additionally, an alternative source for finding company names and identifiers is the Global Legal Entity Identifier Foundation (GLEIF) dataset that provides company names and Legal Entity Identifiers (LEI).

Appendix A Data

A.1 Hazard Data

Storms: We compute the exposure of companies to storms in Europe using the storm footprints from the Climate Data Store, provided by the Copernicus Programme.[6] The data set provides climatological indicators on European storms and their economic impact, derived from the ERA5 reanalysis. We focus on storm footprints, defined as the maximum 3-second 10-m wind gust speed in meters per second (m/s) over a 72-hour period at each model grid point for significant storms. Thus, a storm footprint shows the spatial distribution of maximum wind gust speed for a storm that crosses the area of interest.

TU Delft floods: We calculate the exposures of local units and groups to river and coastal floods using the probabilities of occurrence and return periods developed in Paprotny et al. (2017). The maps are the result of a Bayesian network-based model which generates return period flow rates in European rivers with a catchment area larger than 100km2 with a simulation approach. The simulations are performed using a one-dimensional steady-state hydraulic model, and the results are post-processed using Geographical Information System (GIS) software in order to derive flood zones. The approach of the paper is validated by comparison with the Pan-European map of the Joint Research Center and five local flood studies from different countries. In general, the two approaches show similar results in recreating flood zones from local maps. The simplified approach achieved a similar level of accuracy while substantially reducing computational time. The data also present aggregated results on flood hazards in Europe, including future projections.

A.2 Register of Institutions and Affiliates (RIAD)

RIAD is the ESCB's shared dataset of reference data on individual entities and the relationships between them, as established by EUR-Lex - 02018O0016-20191002 - EN - EUR-Lex (europa.eu). RIAD facilitates the integration of several ESCB databases, namely the CSDB (Centralised Securities Database), the SHSDB (Security Holding Statistics Database) and AnaCredit (Analytical Credit dataset containing detailed information on individual bank loans in the euro area), with consistent entity information. The dataset has been building from 2018 onward and consist of all the entities present in the reporting of the AnaCredit data. The dataset contains over 18mln unique entities, mainly within the eurozone. The information of the entities is enriched from the national database of the eurozone entities.

A.3 E-PRTR

We derive information on the locations of facilities from the E-PRTR. The E-PRTR, as defined in Article 1 of the European Pollutant Release and Transfer Register (E-PRTR), is an integrated pollutant release and transfer register at the community level [...] in the form of a publicly accessible electronic database. It establishes rules for the implementation of the UNECE Protocol on Pollutant Release and Transfer Register, the facilitation of public participation in environmental decision making, and its contribution to preventing and reducing environmental pollution.

According to Article 5 of the E-PRTR Regulation, all operators of facilities undertaking one or more activities listed in Annex I of the Regulation must report specific information if they exceed predefined capacity thresholds. These activities cover sectors such as energy; metal production and processing; the mineral and chemical industries; waste and wastewater management; paper and wood production; intensive livestock production and aquaculture;

[6] More information can be found under Copernicus Program.

food and beverage production; and other industrial activities. As a result, many companies are required to report the locations of their facilities.

The E-PRTR is a public inventory of data submitted by facilities, including information on toxic chemical releases into the air, water, or land; recycling; energy recovery; and off-site transfers for treatment or disposal. A key application of PRTRs is to support sustainability-related decisions by providing insight into facility-level operations over time. Since 2007, the E-PRTR has expanded significantly and now comprises approximately 94,000 facilities across Europe. The E-PRTR provides facility locations, ownership details, and amounts of waste production. Through our data merge, we link approximately 19,049 unique facility owners and 27,022 unique facilities to RIAD.

The E-PRTR covers various industrial sectors but does not encompass all facilities for every company within these sectors. Facility operators are required to report waste production if their facility's output exceeds predefined capacity thresholds. For example, ferrous metal foundries must report waste production if their capacity exceeds 20 tons per day. However, for some industries, there is no capacity threshold.[7]

[7] More details on general applications of PRTRs and reporting requirements are available in (EPA, 2006; OECD, 2017)

Appendix B Fuzzy name matching

B.1 Data cleaning

In the pre-processing stage, the main goal is to reduce the variability in company names, which is critical for the accuracy of the subsequent name matching steps. The first step involves minimizing the number of unique characters by transliterating non-ASCII characters, removing punctuation, and converting all characters to lowercase (see Figure 2.1). This reduces the character set from more than 300 distinct characters to around 40.

In addition, company names often include acronyms or legal suffixes indicating the type of entity or generic terms commonly found in many names. These legal suffixes are removed to prevent any impact on matching algorithms, since companies that only differ by the legal suffix are most commonly all parts of the same corporate group. Common terms like "International" are abbreviated or omitted, minimizing their influence in string similarity measures.

The process of standardizing company names in E-PRTR and RIAD leads to a reduction of unique company names. For instance, 'Company Name # Ltd' and 'Company Name GmbH' are identical after preprocessing, namely *company name*. This can lead to an m:n-relationship of matches on the original company names. To limit the number of duplicate matches after preprocessing, we implement a waterfall approach to identify 'perfect matches', i.e. company names that are identical after preprocessing. In every preprocessing step the perfect matches between E-PRTR and RIAD are identified and the matched company names are not taken into account in later preprocessing steps. The preprocessing steps are as follows:

1. **preprocessing step 1:** all names in lowercase, removal of special characters, transliteration non-ASCII characters;
2. **preprocessing step 2:** abbreviation of all legal prefixes and suffixes, for instance 'limited' to 'ltd';
3. **preprocessing step 3:** removal of all legal prefixes and suffixes;
4. **preprocessing step 4:** removal or abbreviation common terms.

The above waterfall approach leads to the identification of 22,469 perfect matches, of which 14,351 in **step 1**, 4,011 in **step 2**, 3,077 in **step 3** and 1,030 in **step 4**. Hence, most perfect matches are identified after minimal preprocessing (steps 1 and 2). However, not all perfect matches are meaningful. For instance, preprocessing step 4 can result in empty names, such as '???GmbH' to '', or names that only consist of an abbreviation, such as 'International Ltd' to '#int#'. We apply expert judgement to remove non-meaningful perfect matches and are left with a total of 25,541 E-PRTR facilities matched to 17,932 RIAD companies. After removing the perfect matches from the preprocessed E-PRTR company names, we are left with 23,477 unique E-PRTR company names and 11,993,912 unique RIAD company names. For these remaining E-PRTR company names we apply the fuzzy name matching to identify non-perfect matches.

B.2 Calculation of similarity measures

Van Der Loo (2014) categorizes string distance algorithms into three types: edit-based distances, heuristic distances, and token-based distances, each addressing different aspects of string comparison. For each of these three categories, we select one similarity measure.

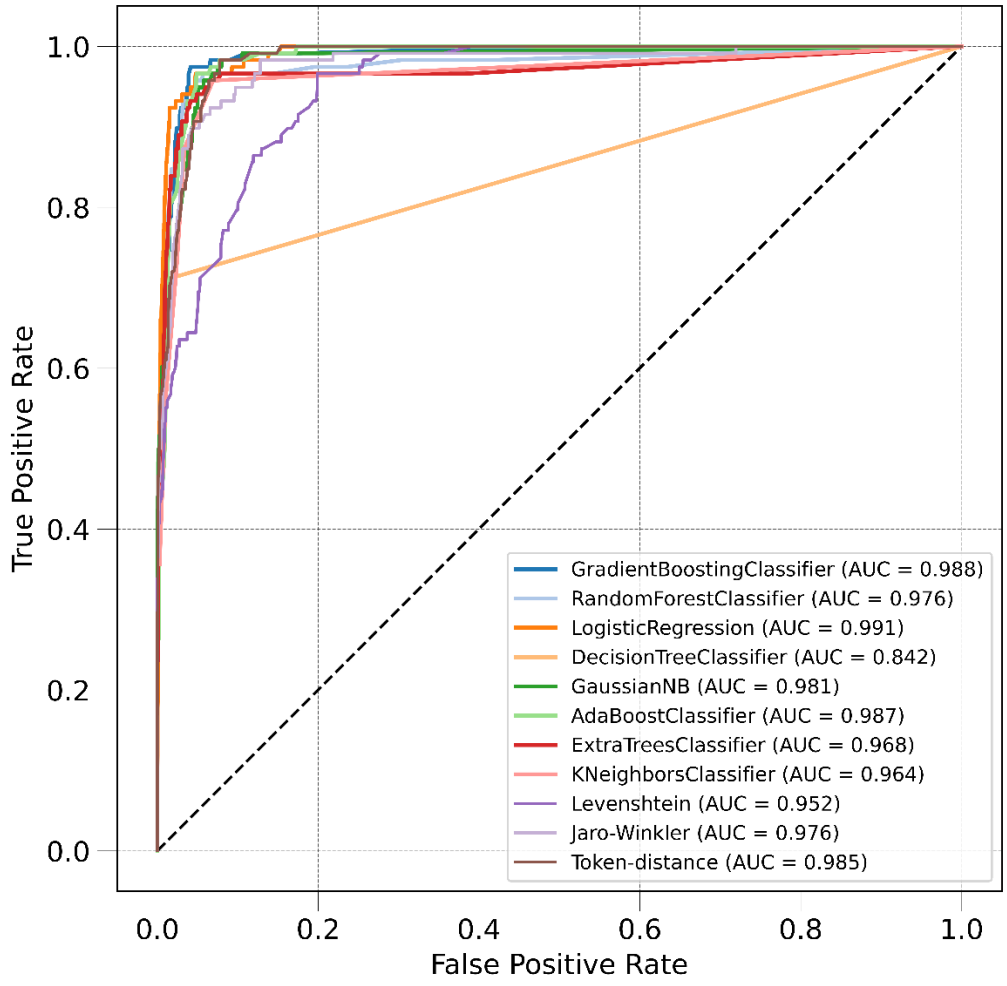
In our paper, the first similarity measure we use is the Levenshtein distance, which falls into the edit-based category. It assigns a uniform cost of one unit for each operation, be it insertion, deletion, or substitution (Aouragh et al., 2015). The second metric that we examine is the Jaro-Winkler similarity, which belongs to the

heuristic distance category. This measure calculates the shared characters between two strings, making it particularly effective for shorter strings such as company names. A modification of the Jaro distance, it emphasizes matching prefixes, which is especially useful in contexts where prefix alignment is crucial for identifying matching company names (Gali et al., 2019). Incorporating prefix length implies that matching prefixes suggest a higher likelihood of relatedness, resulting in a higher similarity score. Finally, the Q-gram distance measures string similarity by comparing substrings of fixed length (q), such as bigrams (2 characters) or trigrams (3 characters). It counts the number of q-grams that do not match between two strings and divides this over the number of q-grams in the strings, with a count of 0 indicating a perfect match and a score of 1 indicating no overlap whatsoever (Denman et al., 2019).

B.3 Model selection and performance

For every preprocessed name in E-PRTR 50 potential matches are selected from RIAD based on the cosine similarity of TF-IDF transformed trigrams. For each of the 50 potential matches, the chosen similarity measures are calculated. The similarity measures used for this analysis are the Levenshtein distance, the Jaro-Winkler distance, and the 3-gram distance. Next, we manually annotate 5% of the names in E-PRTR. To ensure enough actual matches, we only annotate company names for which at least one potential match has at least one similarity measure with a score of 50% or more. Of the annotated data, 240 company names in E-PRTR have a match in RIAD and 904 company names in E-PRTR do not have a match in RIAD. The resulting training and test set therefore contains 240 matches and $240 * 49 + 904 * 50 = 56{,}960$ nonmatches. The annotated data are divided 70%-30% into a training set and a test set. The training set is balanced by selecting for every match one nonmatch. Next, several supervised machine learning algorithms are trained on the training set and applied to the unbalanced test set. The machine learning models under consideration are Gradient Boost (GB), Random Forest (RF), logistic regression (Logit), Decision Tree (Tree), Naive Bayes (NB), AdaBoost (AB), Extra Tree (ET) and k-nearest neighbor (KNN). Figure B.1 shows the receiver-operating characteristic curve of the models. The figure shows that the Gradient Boost, logistic regression and AdaBoost perform better than the best performing single similarity measure (3-gram distance). The other algorithms do not outperform the 3-gram distance.

Figure B.1 Receiver-operating characteristic curve of the classification models[8]



Next, the outcomes are cross-validated using the combined training and test set. The cross-validation is performed over five folds, where for every fold the annotated data are divided into a balanced training set and an unbalanced test set. Table B.1 shows the mean and the standard deviation of several classification performance measures over the five folds using a classification threshold of 0.9. The highest precision (true positives divided by true positives plus false positives) is obtained for the logit model at 90.8%, while still maintaining high accuracy (98.6%). Based on the model performance we select the logistic regression with a threshold of 0.9 to identify additional matches. This results in an additional 1,481 E-PRTR facilities being matched to 1,223 RIAD entities.

[8] The figure shows the ROC curve of the classification models tested to match company names. Additionally, the figure shows the ROC curve of the individual similarity measures underlying the classification models, namely Levenshtein, Jaro-Winkler and Q-gram. The ROC curve shows the relationship between the false positive rate and true positive rate for different thresholds on the unbalanced test set. The Area Under the Curve (AUC) shows how close the performance of a given machine learning model is to that of a model that perfectly classifies matches and nonmatches (AUC of 1).

Table B.1 Supervised machine learning algorithms by classification performance[9]

		GB	RF	Logit	Tree	NB	AB	ET	KNN
Accuracy	μ	98.5	98.5	98.6	90.2	98.8	98.6	98.6	98.5
	σ	0.2	0.0	0.1	13.0	0.0	0.1	0.1	0.1
F1	μ	58.9	48.2	49.4	38.2	0.0	53.0	53.0	46.9
	σ	5.1	0.9	6.1	16.2	0.0	2.1	2.1	4.5
Precision	μ	63.8	83.9	90.8	29.1	0.0	79.0	79.0	80.8
	σ	6.1	4.6	1.6	14.2	0.0	3.4	3.4	7.7
Recall	μ	54.7	33.9	34.3	69.5	0.0	39.9	39.9	33.2
	σ	4.5	0.9	6.0	5.1	0.0	1.8	1.8	4.0

[9] The table reports the accuracy, F1 ratio, precision and recall for all classification models tested to match company names over the cross-validation samples. The cross-validation was performed on the combined training and test set using five folds. For every fold the annotated data was divided into a training set and a test set, after which the training set was balanced. The reported results are for a threshold of 0.9. The methods presented include Gradient Boost (GB), Random Forest (RF), logistic regression (Logit), Decision Tree (Tree), Naive Bayes (NB), AdaBoost (AB), Extra Tree (ET) and k-nearest neighbor (KNN).

References

Aouragh, S. L., Gueddah, H., Yousfi, A., Sidi, U., Abdellah, M. Ben, & Hicham, G. (2015). Adaptating the Levenshtein Distance to Contextual Spelling Correction. *International Journal of Computer Science and Applications, Technomathematics Research Foundation, 12*(1), 127–133. https://www.researchgate.net/publication/273758433

Atlan, A. L., Borea, G., Mateo, C., Graciano, C., Kaissoumi, L. El, Gosset, L., Osiewicz, M., Pio, L. Y., & Py, L. (2024). Location of physical assets-addressing one of the main data gaps in assessment of climate-related risks. *11th European Conference on Quality in Official Statistics*. https://www.cgfi.ac.uk/spatial-finance-initiative/geoasset-project/geoasset-databases/

Bauer, R., Broeders, D., & De Carolis, F. (2025). Local institutional ownership and price discovery around extreme weather events. *ECB Working Paper Series*, *3069*. https://doi.org/10.2866/8995136

Braun, A., Braun, J., & Weigert, F. (2024). *Extreme Weather Risk and the Cross-Section of Stock Returns*. https://ssrn.com/abstract=3952620

Bressan, G., Duranovic, A., Monasterolo, I., & Battiston, S. (2024). Asset-level assessment of climate physical risk matters for adaptation finance. *Nat Commun*, *15*(5371), 1–38. https://doi.org/10.2139/ssrn.4062275

Christiaen, C., Jackman, A., & Lockwood, P. (2025). Location, Location , Location: Asset location data sources for nature-related risk analysis. *Spatial Finance Initiative*, 1–25.

Denman, K., Fortier-Shultz, V., & Maus, A. (2019). *Assessing Record Linkage Matches Using String Distance Measures*. https://en.wikipedia.org/wiki/Damerau–Levenshtein_distance

ECB. (2023). Towards climate-related statistical indicators. *ECB, Available at: Https://Www.Ecb.Europa.Eu/Stats/All-Key-Statistics/Horizontal-Indicators/Sustainability-Indicators/Html/Index.En.Html (Accessed 11 March 2025)*, *January*. https://www.ecb.europa.eu/pub/pdf/other/ecb.climate_change_indicators202301~47c4bbbc92.en.pdf

ECB. (2024). Climate change-related statistical indicators. *ECB, Available at:"https://Www.Ecb.Europa.Eu/Pub/Pdf/Scpsps/Ecb.Sps48~e3fd21dd5a.En.Pdf"*, *48*. https://doi.org/10.2866/059096

EIOPA. (2022). European insurers' exposure to physical climate change risk - Potential implications for non-life business. *EIOPA Discussion Paper*, *22/278*(May).

EPA. (2006). Guidance Document for the implementation of the European PRTR. *European Commission, Available at: Https://Www.Epa.Ie/Publications/Compliance--Enforcement/Licensees/Reporting/European-Commission-Guidance-Document-for-the-Implementation-of-the-e-Prtr-.Php (Accessed 11 March 2025)*.

European Environment Agency. (2023). European Union 8th Environment Action Programme: Monitoring report on progress towards the 8th Environment Action Programme objectives. *EEA Report*, *11*. https://doi.org/10.2800/34224

Gali, N., Mariescu-Istodor, R., Hostettler, D., & Fränti, P. (2019). Framework for syntactic string similarity measures. *Expert Systems with Applications*, *129*, 169–185. https://doi.org/10.1016/j.eswa.2019.03.048

Germeshausen, Robert; Chlond, Bettina; Tchorzewska, Kinga; von Graevenitz, K. (2022). ME-FINE: Mannheim European panel on Financial Indicators and Emissions. *ZEW Dokumentation, ZEW - Leibniz-Zentrum Für Europäische Wirtschaftsforschung*, *22–01*.

Hong, H., Li, F. W., & Xu, J. (2019). Climate risks and market efficiency. *Journal of Econometrics*, *208*(1), 265–281. https://doi.org/10.1016/j.jeconom.2018.09.015

Huizinga, J., de Moel, H., & Szewczyk, W. (2017). Global flood depth-damage functions. Methodology and the database with guidelines. *Joint Research Centre (JRC)*, *EUR 28552*(10.2760/16510). https://doi.org/10.2760/16510

Loberto, M., & Russo, R. (2024). Climate risks and firms: a new methodology for assessing physical risk. *Mimeo*.

Nijhuis, M. (2022). Company Name Matching. *Medium, Available at: Https://Medium.Com/Dnb-Data-Science-Hub/Company-Name-Matching-6a6330710334 (Accessed 22 May 2025)*. https://medium.com/dnb-data-science-hub/company-name-matching-6a6330710334

OECD. (2017). Framework on the role of pollutant release and transfer registers (PRTRs) in global sustainability analyses. *OECD Environment Directorate, Available at: Https://Prtr.Unitar.Org/Site/Document/1313 (Accessed 11 March 2025)*.

Paprotny, D., Morales-Nápoles, O., & Jonkman, S. N. (2017). Efficient pan-European river flood hazard modelling through a combination of statistical and physical models. *Natural Hazards and Earth System Sciences*, *17*(7), 1267–1283. https://doi.org/10.5194/nhess-17-1267-2017

Van Der Loo, M. P. J. (2014). The stringdist Package for Approximate String Matching. *The RJournal*, *1*.