BRUEGEL WORKING PAPER 2014/10ii

A SCALEABLE APPROACH TO EMISSIONS-INNOVATION RECORD LINKAGE

MARK HUBERTY*, AMMA SERWAAH** AND GEORG ZACHMANN†

Highlights

- This paper reports an approach to linking data on European emitters to data on their innovation practices. We illustrate a straightforward approach to record linkage between the European Union Community Integrated Transaction Log (CITL) and the PATSTAT international patent database. We show how that record linkage can be maintained with relatively minimal human input.

* University of California, Berkeley; markhuberty@berkeley.edu
** Bruegel
† Bruegel; georg.zachmann@bruegel.org

This work was supported by the SIMPATIC project (EU Seventh European Union Framework Programme, grant agreement no. 290597). Mark Huberty received additional support from the United States Environmental Protection Agency STAR fellowship for work relating to the political economy of green innovation. We thank in particular Lee Fleming for use of his disambiguation engine and helpful consultation; Forest Gregg and Derek Elder, who wrote the dedupe library for Python; and KU Leuven, for providing the benchmark data based on their disambiguation of PATSTAT. All errors remain our own.

SEPTEMBER 2014

A scaleable approach to emissions-innovation record linkage*
Working Paper

Mark Huberty[†], Amma Serwaah[‡], and Georg Zachmann[§]

June 30, 2013

Abstract

This paper reports an approach to linking data on European emitters to data on their innovation practices. We illustrate a straightforward approach to record linkage between the European Union Community Integrated Transaction Log (CITL) and the PATSTAT international patent database. We show how that record linkage can be maintained with relatively minimal human input.

1 Introduction

The European Union Emissions Trading Scheme (ETS) fulfills several functions. Innovation is central to those functions. The emissions price generated by the ETS should theoretically incentivize firms to invest in low-emissions technologies. Given a sufficiently

[*] This work was supported by the SIMPATIC project under the Seventh European Union Framework Programme. Mark Huberty received additional support from the United States Environmental Protection Agency STAR fellowship for work relating to the political economy of green innovation. We thank in particular Lee Fleming, Forest Gregg, and Derek Elder for their contributions to disambiguation approaches. All errors remain our own.

[†] University of California, Berkeley. markhuberty@berkeley.edu

[‡] Bruegel.

[§] Bruegel. georg.zachmann@bruegel.org

credible long-term emissions price, those investments should fulfill not just today's needs for marginal improvements to existing technologies, but also the long-term need for transformative innovation to serve the needs of a low-carbon economy.

Estimating whether this has occurred is complicated by several data problems. First, innovation data are readily available, but they pose significant data quality barriers. Second, linking patterns of innovation to exposure to the emissions price itself requires a means of linking data on innovative activity to data on exposure to the European emissions price. Given the very large number of European firms relative to the number of firms regulated under the ETS, establishing these links amounts to looking for a needle in a haystack. Finally, many firms may separate business units most exposed to the ETS permit price from those most directly engaged in innovation. Hence any linkage strategy must capture both direct and subsidiary connections, lest it underestimate the innovation effect of the EU ETS.

This paper describes an approach to linking innovation and emissions data for the European Union. We (Huberty et al., 2013) have previously introduced a flexible, scaleable approach to disambiguation of the PATSTAT database on international patenting activity. This paper builds on that result, providing a machine-learning based approach to record linkage between the disambiguated PATSTAT data and firms regulated by the European Union Emissions Trading Scheme. We illustrate that this approach can achieve relatively high levels of accuracy with minimal human input.

2 Linking innovation and regulation

We implement a three-stage process for matching firms regulated under the European Emissions Trading Scheme to their innovation activities as they appear in patent data. We first describe a means of establishing a pool of likely matches. We then demonstrate a supervised learning approach to filtering these likely matches. We provide estimates that this approach is highly accurate compared with manually-labeled data. Finally, we provide summary statistics showing, in broad terms, how the pattern of identifiable links between regulated and innovating firms breaks down across European member states and regulated sectors.

2.1 Data

We draw on two different data sets for the purposes of record linkage. First, as described in Huberty et al. (2013), we disambiguate the October 2011 version of the PATSTAT database for each of the EU-25 countries (excluding Cyprus and Malta). This results in a one-to-many map between unique individuals and their non-unique occurrences in PATSTAT. For each unique individual, we consolidate this map to a single record comprised of the most common name variant and the most common non-null latitude and longitude pair.

Second, we use the most recent version of the Community Integrated Transaction Log (CITL), which tracks regulated firms and emissions permit allocation under the European Emissions Trading Scheme. The CITL data represent an official accounting log; hence we assume with confidence that each account holder is a unique entity, and do not disambiguate further. Account holders were geo-coded using a fuzzy geo-coding algorithm as described in Huberty et al. (2013). Both data sets were, to the extent possible, standardized for case, punctuation, and use of name diacritics.

2.2 Record linkage

We use record linkage here to refer to a specific form of data disambiguation: given two records from two different databases, identify whether they refer to the same entity. This problem is related to, but separate from, our earlier work (Huberty et al., 2013) focused on disambiguation *within* a single database. In that instance, the primary problem was twofold: first, how to find all versions of the same entity given possibly significant variance in the format and spelling of inventor names; and second, how to do so with minimal human intervention and in finite time. We show that the approach described by Bilenko (2006) and implemented in the `dedupe` library for Python performed well on both counts.

Matching between the disambiguated patent database and the CITL account holders' database poses a variant on this problem. The mismatch between databases in terms of data availability complicates matters by reducing available data with which to compare records. Even with the data quality issues discussed in Huberty et al. (2013), internal comparisons within PATSTAT could draw on (at a minimum) name, geography, innovation categories, and coauthors to distinguish one inventor from another. In contrast, matching PATSTAT to CITL can only draw on name and geography.

Conversely, cross-database matching simplifies the problem by greatly restricting the set of matches we need to examine. Whereas in the PATSTAT database the comparison set scaled as N^2 for N records in the PATSTAT data set, the PATSTAT-CITL linkage requires at most $N \times M$ comparison for M CITL records. Since there are only approximately ten thousand registered emitters in the Emissions Trading Scheme, $M \ll N$. This feature greatly diminishes the need for sophisticated blocking strategies to keep computation tractable.

To match records *across* databases, we implement the following algorithm. In each case, the PATSTAT data refers to the output of the `dedupe` process discussed in Huberty et al. (2013). That output consolidated all variants on a unique inventor to a single record by taking (1) the most common name variant and (2) the most common non-null latitude/longitude pair:

1. For each account holder in the CITL database:

    (a) Select the PATSTAT records filed in the country of origin for that account holder

    (b) Compute the Levenshtein ratio of the CITL name and each PATSTAT name

    (c) Return the top $N = 3$ closest matches.

2. For each potential CITL:PATSTAT match:

    (a) Compute multiple variants of name similarity between the PATSTAT and CITL names. We used the Levenshtein edit ratio; and the Jaccard unigram similarity metric

    (b) Compute geographic distance has the Haversine great circle distance between record pairs[1]

3. Label a subset of all possible matches as either match (1) or non-match (0), in two steps:

    (a) Compute an initial set of matches by taking only PATSTAT:CITL pairs with exact name matches

    (b) Expand the set of matches and non-matches by hand-labeling a random subset of pairs from the remaining data. Labeling continued until non-matches numbered 25% of the exact match volume.

4. Train a support vector classifier on the labeled data, using as predictor variables the two name distances (Levenshtein and Jaccard) and the geographic distance[2]

5. Use the classifier to predict matches and non-matches for the entire set of potential matches

[1] Each CITL account holder address was geo-coded using a city-level geo-coding algorithm. For more details on the algorithm, see the `fuzzygeo` package at https://github.com/markhuberty/fuzzygeo.

[2] All classification and estimation used the `sklearn` library for python.

Classification accuracy was estimated at 92±1% via ten-fold cross-validation. We note several caveats for using this estimation strategy. First, the labeling rubric should be fairly generous towards classifying different country subsidiaries together, in order to capture instances where corporate structures assign patents to a research and development division separate from the operations division responsible for ETS-regulated activities. Second, even this more generous labeling strategy will not account for situations in which CITL emitters have wholly-owned subsidiaries with completely different names, who may be involved in innovation activities. Hence the matches reported here represent, at best, a conservative estimate of the true web of connections between innovation activity and direct exposure to emissions pricing. Third, many emitting firms may collaborate with suppliers on innovation, even if they are not listed as coauthors on patents themselves. Hence, for instance, Siemens may conduct significant research and development in collaboration with its wind turbine customers, but those links may not show up in the patents filed to protect the results of that R&D. These issues–some of which go well beyond the problem of identifying likely matches between CITL and PATSTAT entities–may therefore understate the scope of the entities affected by ETS regulation.

We note that the classifier, once trained, should be useful for some period into the future. Having trained the classifier on thousands of potential record pairs, the relative degree of additional variance introduced by either new CITL account holders or new inventors in the PATSTAT database should be low. Hence the classification approach presented here should permit ongoing updates to the record linkage estimates without substantial further investment in human coding effort.[3]

[3] We are pursuing the use of additional data in the classifier. In particular, we hope to develop a means of associating inventors to economic sectors based on the technology categories in which they innovate. Since the CITL data have economic sector as well, this would permit matching not only to name and place, but economic activity as well. This would help distinguish, in particular, name differences due to minor spelling errors from name differences due to real distinctions among firms. This work is presently ongoing.

2.3 Results and descriptive statistics

We present three measures of match coverage. First, as figure 1 shows, member states' proportion of CITL account holders with positive PATSTAT matches varied from zero percent at the low end (for Latvia and Lithuania) to over forty percent for the Netherlands and Italy. Second, as table 1 shows, match rates varied widely across economic sectors. Bulk chemicals firms and steel manufacturers were the most likely to have records in the innovation database. In contrast, airlines and small power plant operators were least likely. Finally, figure 2 shows that these match rates break down by the CITL installation classifications.

These results comport with a fairly simple prior expectation for innovation across different economic sectors. Large industrial chemical and specialty metallurgy firms invest in a variety patentable of process and product innovations. In contrast, while airlines may identify how new technologies might help serve new markets, they nevertheless buy most of that technology from a handful of major international suppliers, while focusing their internal efforts on operational efficiency.[4] Hence we would expect substantial variation in PATSTAT:CITL match rates both within countries (consequence of sectoral variation in innovation activity) and between countries (consequence of variance in industrial specialization and innovative activity).

3 Conclusions

We have demonstrated a relatively straightforward method of tying firm innovation activities to their regulatory exposure in the European Emissions Trading Scheme. We demon-

[4] See here, for instance, Brueckner and Pai (2009), who note that most capital innovation in the airline industry comes from suppliers, rather than operators; while operational innovations derive from new technological capabilities.

Sector	Pct. Match to PATSTAT
Aircraft	0.10
Coke ovens	0.14
Combustion, > 20MW	0.26
Combustion	0.00
Pulp and paper	0.21
Raw ceramics and brick	0.25
Glass and glass fibre	0.33
Cement clinker	0.31
Pig iron and steel	0.28
Ceramic products	0.42
Metal ore	0.13
Oil refineries	0.37
Other	0.29
Bulk chemicals	0.50
Carbon black	0.00
Ferrous metals	0.50
Gypsum and plaster	0

Table 1: Percentatge of CITL installations by sector with positive PATSTAT matches. Results shown for the EU-25, Cyprus and Malta omitted

strate that this method performs well against manually-checked records. We suggest, however, that this effort remains incomplete given the likelihood of complex corporate structures that are not amenable to record linkage based on comparisons of name and geography alone. Hence the results presented here provide a conservative estimate of the degree of innovative activity among regulated firms. We propose, and are pursuing, additional data on economic sector and operations that may assist in improving these matches.

References

Bilenko, M. Y. (2006). *Learnable similarity functions and their application to record linkage and clustering*, volume 67.

Brueckner, J. K. and Pai, V. (2009). Technological innovation in the airline industry: The impact of regional jets. *International Journal of Industrial Organization*, 27(1):110–120.

Huberty, M., Serwaah, A., and Zachmann, G. (2013). A flexible, scaleable approach to the international patent “name game”. Working paper, Bruegel.

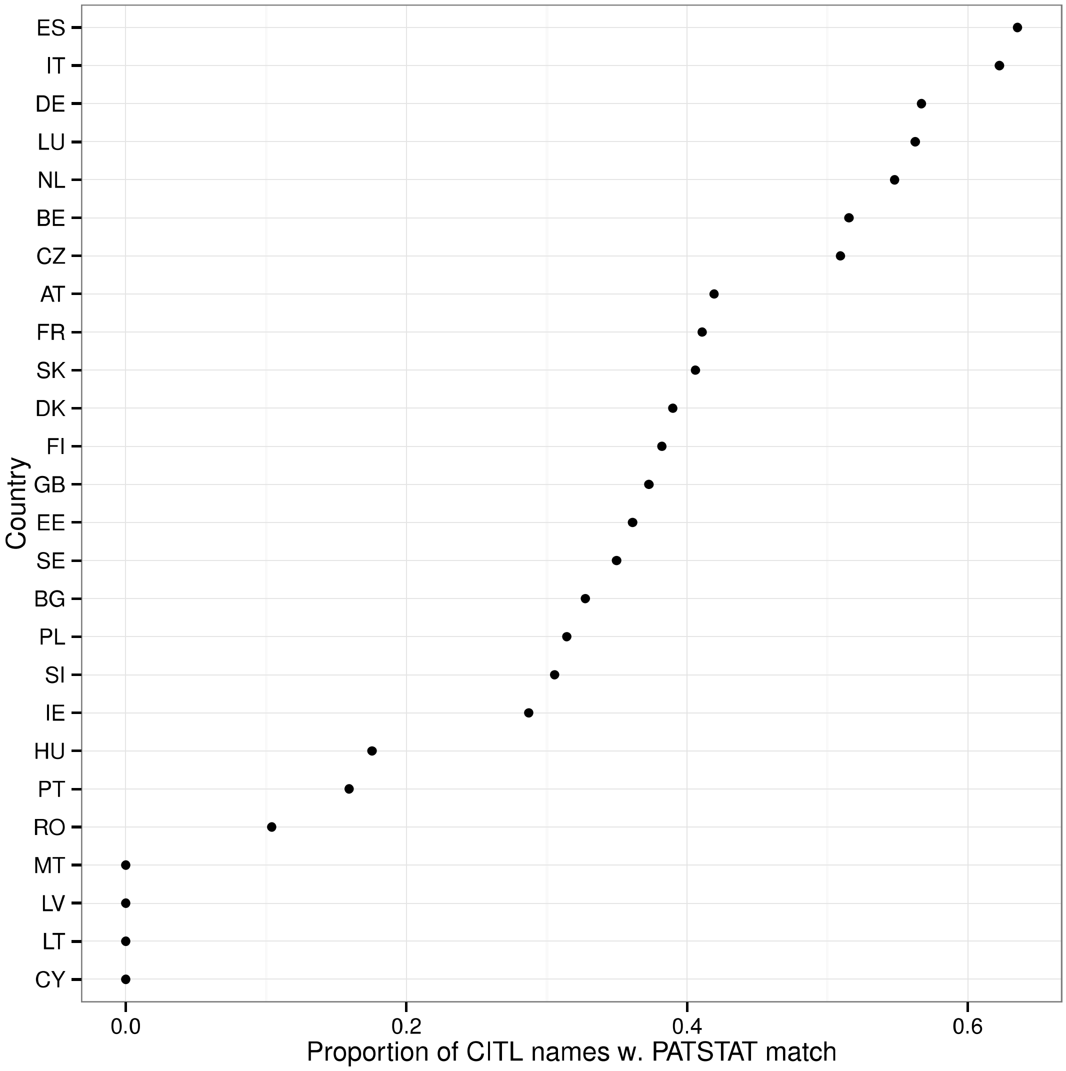

Figure 1: Proportion of CITL account holders with positive PATSTAT match. This figure illustrates the overall proportion of CITL account holders in each country for which a positive match to a PATSTAT inventor was found.

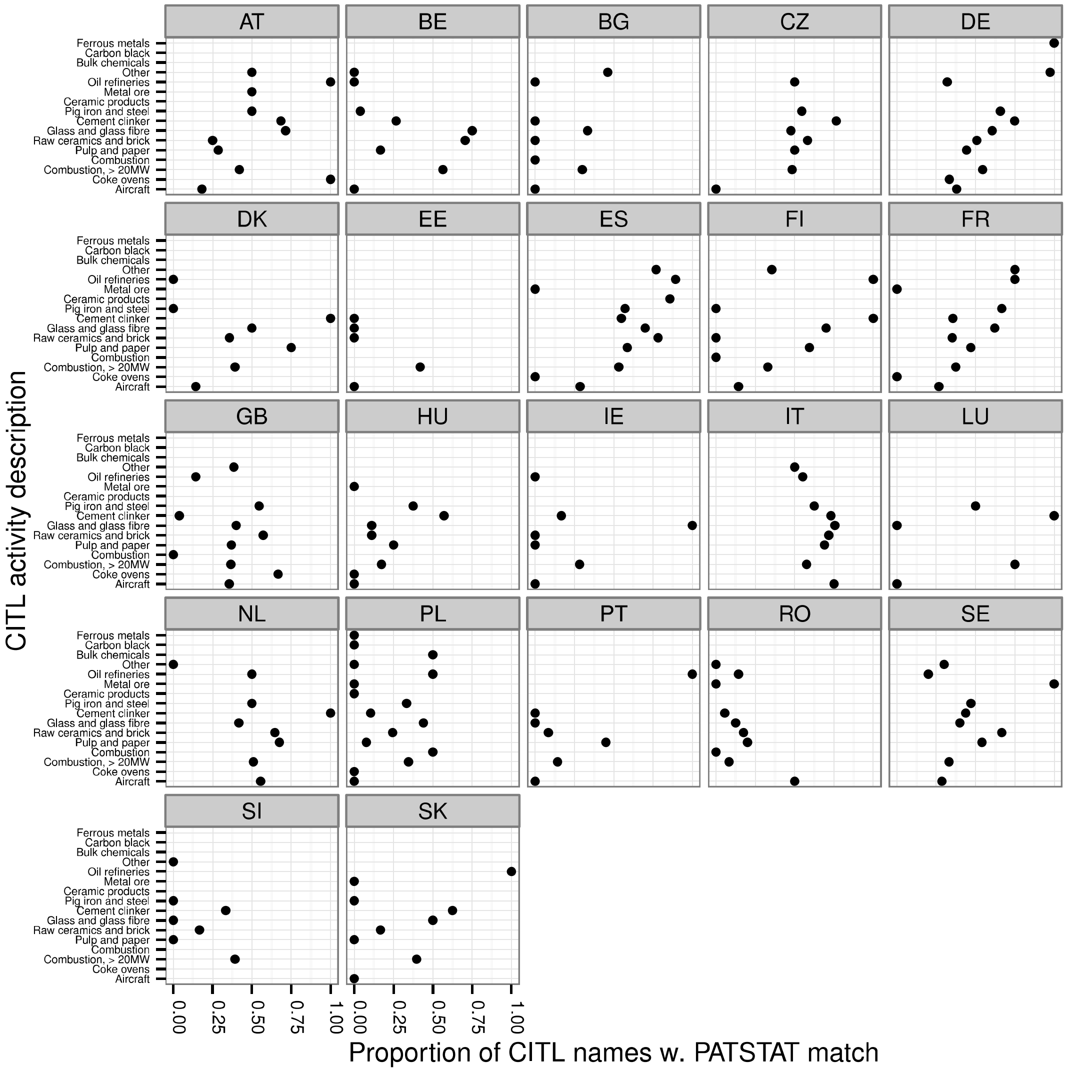

Figure 2: Proportion of CITL account holders with positive PATSTAT matches, by country and installation type. This figure illustrates the overall proportion of CITL account holders in each country for which a positive match to a PATSTAT inventor was found, broken out by the type of installation. Installation types correspond to the CITL installation type registry.