Material and immaterial regional interdependencies: using the web to predict regional trade flows

Emmanouil Tranos
e.tranos@bristol.ac.uk
University of Bristol and The Alan Turing Institute
UK

Andre Carrascal Incera
carrascalandre@uniovi.es
University of Oviedo
Spain

George Willis
GCW519@student.bham.ac.uk
University of Birmingham
UK

ABSTRACT

This paper brings new web data and machine learning methods in order to expose the structure and the evolution of regional trade interdependencies. Although our focus is on the UK regions, the proposed research framework lends itself to applications to other countries. Interregional trade relationships have traditionally been very difficult to capture, as national statistics do not monitor intra-country trade links at the firm level. Recently, the EUREGIO project (Thissen et al., 2018) has published spatially disaggregated Input-Output (I-O) information for 37 NUTS2 UK regions for 14 industries, which includes imports and exports by region of origin and destination. This database was developed using the interregional trade data from the PBL Netherlands Environmental Assessment Agency, freight transport data from Eurostat (for goods), and business flight ticket information (for services). Nevertheless, building such an interregional trade dataset has been a costly and non-trivial process. In this paper, we are proposing a new research framework to predict such flows by utilising freely available web data. Specifically, we are employing the JISC UK Web Domain Dataset in order to extract hyperlinks between geolocated commercial websites in the UK. This dataset is a subset of the Internet Archive, which includes all the archived webpages under the .uk country code Top Level Domain (ccTLD). We are able to geolocate these webpages by searching the archived web text for the inclusion of a UK postcode. In addition, this dataset also contains the hyperlinks included in the archived webpages and we are able to aggregate these data and create an interregional network based on the hyperlinks between geolocated commercial webpages. Formally, we approach the prediction of the trade flows as a regression problem. Hence, we employ some well-established machine learning models, such as Random Forests, to predict interregional trade flows using, among other features, the network of digital interdependencies between the UK regions. Our results indicate a very high capacity of the interregional hyperlinks feature to predict interregional trade indicating that trade leaves behind significant digital breadcrumbs.

CCS CONCEPTS

• **Information systems** → **Web mining**; • **Applied computing** → **Digital libraries and archives**.

KEYWORDS

interregional trade, web archives, web data, machine learning, prediction, random forests

ACM Reference Format:
Emmanouil Tranos, Andre Carrascal Incera, and George Willis. 2018. Material and immaterial regional interdependencies: using the web to predict regional trade flows. In *WebSci '21: ACM Web Science Conference, June 21–25, 2021, University of Southampton, UK*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Bilateral trade is a complex phenomenon per se [67], but its complexity increases when it is approached from a spatially disaggregated perspective. Regions[1] behave different from countries from an economic perspective as they are more specialised in specific sectors and more open to trade with other regions in comparison to national economies. This makes them face very important external dependencies. Also, regions vary greatly in terms of their specialisation patterns, and therefore, there is great variation in terms of trade relationships and openness within regions. Furthermore, because of globalisation patterns and the spatial fragmentation of production, regional trade of intermediate and final products is no longer constrained to interregional transactions within countries. So, by lowering trade barriers in the past decades, restrictions to international trade declined and the external dependence of regions became global. Events happening in regions of countries from the other side of the globe can affect closer regions through disruptions in Global Supply Chains, as the Covid-19 crisis is showing us. A region would be affected by an economic downturn in a second region if it sells much of its production to that region (directly), or if it sells its production to regions that sell their production to that region (indirectly), while regions less dependent on that second region might be hurt to a much lesser extent when in crisis [75]. This is why, among other factors, regions had significantly divergent experiences in avoiding or overcoming economic shocks [38].

Consequently, understanding and, if possible, predicting regional trade is key to comprehend regional economic performance and the exposure to internal and external shocks, but also to articulate proper place-based development policies [5]. Interregional relations and modern supply chains are central in a systemic way of thinking

[1] In the paper we shall use the terms 'regional' and 'subnational' interchangeably.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WebSci '21, June 21–25, 2021, University of Southampton, UK
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

about regional innovation and growth strategies [77] such as the smart specialisation policy initiatives [52].

The big caveat is the lack of sectoral, interregional trade data, which are absent from key cross-country data providers such as the Eurostat and OECD. One exception is the work of Thissen et al. [77], who followed the parameter-free Simini et al. [68] approach and estimated interregional trade flows between 256 European NUTS2 regions at a sector level by disaggregating national input-output tables. These data have been utilised in regional economics research – see discussion and references in Section 4 – and are nowadays the go-to interregional trade data set. Nevertheless, production of such data are neither simple nor easily reproduced.

Our paper contributes to this line of inquiry by utilising state-of-the-art machine learning algorithms and novel web data to make out-of-sample predictions for the UK NUTS2 regions during the period 2000-2010. Specifically, we use open and archived web data to create counts of hyperlinks between websites that we are able to geolocate. We feed such variables to a Random Forest (RF) model, alongside a limited number of other predictors, and we are able to achieve accuracy scores above 90% in predicting *unseen* interregional trade flows. Our underpinning hypothesis is that trade leaves behind digital breadcrumbs (Rabari and Storper [63]), which can be effectively utilised to predict interregional trade flows, which are both important for regional policies and also very difficult to observe.

Modelling interregional trade flows has traditionally been within the core of geographical research as it well embedded within the discipline's effort to explain the determinants of aggregated interactions across space [for a recent review see 58]. Methodological and conceptual developments on *spatial interaction models* have been extensively employed in order to model flows of trade between regions [11, Paul Lesage and Polasek [60]] and countries [14, 15]. Following current debates within the quantitative geographical thinking [69] and, more broadly, computational social sciences [43], geographical research has been focusing more on explaining the determinants of interregional trade flows than predicting such flows.

This paper is aligned with the state-of-art within the Data Science research domain and its subsequent epistimological effects in geography [13, 69] and economics [39] regarding the role of machine learning algorithms in making out-of-sample predictions of data instead of focusing on explanatory research frameworks. Simply put, the above advocate towards the use of ML algorithms, such as RF, as they outperform ordinary least squares – still one of the widely used estimators to model interregional trade flows – in out-of-sample predictions even when using moderate size training datasets and limited number of predictors [4, 54]. Such an approach can be particularly useful for predicting interregional trade flow given the scarcity and cost to produce such data.

RF have been extensively utilised in answering regression type of problems. Pourebrahim et al. [62] coupled a traditional methodological framework – gravity and spatial interaction modelling – with ML techniques including RF as well as data from online social media to predict commuting flows in New York City. Lima and Delen [46] used ML and RF to predict and explain corruption across countries. Sinha et al. [70] open a dialogue on the need for spatial ensemble learning approaches, such as RF, aimed to be used with spatial data with high autocorrelation and heterogeneity. Credit [13] introduced spatially explicit RF to predict employment density in Los Angeles. Guns and Rousseau [21] use RF to predict and recommend high-potential research collaborations, which have not yet been materialised.
Ren et al. [64] train RF as well as other classifiers to predict socio-economic status or cities using a variety of online and mobility predictors. Wang et al. [87] train RF and other machine learning algorithms to predict urban socio-economic characteristic based on 311 service requests[2]. RF tend to perform better than other algorithms in predicting, for example, real estate prices, income per capita, percentage of residents with a graduate degree, percentage of unemployed residents, percentage of residents living below the poverty level, as well as demographic characteristics for 30 US cities.

The structure of the paper goes as following. The next section reviews the literature which either highlighted the lack of interregional trade data or employed innovative and often data-intensive approaches to capture such flows. Then, we describe the methods and the data we use and present the results of the analysis. The paper ends with a conclusions section.

2 FROM THE LACK OF REGIONAL TRADE DATA TO WEBDATA

The lack of bilateral trade data resulted in a very prolific branch of the literature attempting to estimate trade flows at country and regional level. Without a doubt, the most important step in this regard was the introduction of gravity equations in the early works of Tinbergen [80], Linnemann [48] and Leamer and Stern [44]. In summary, a gravity equation is based on the idea that bilateral trade between two territories depends on their sizes (expressed normally as GDP or GDP per capita) in relation to the distance between them or transport costs (as an impediment factor), and some preference factors (common border, common language, etc.) [1, 18]. In the last years, the emphasis has been placed on discussing the proper estimation methods to accurately predict trade flows (OLS, Tobit, panel fixed effects, Heckman two-step, etc.). A review of the alternative methods applied in gravity models can be seen in Gómez-Herrera [20].

A different strand of the literature comes from the multisectoral trade analysis of Input-Output flows. While the theoretical framework of multiregional Input-Output databases was developed in the 1950s [30], the biggest empirical take-off did not come until the release of the World Trade Organisation databases such as the WIOD (World Input-Output Database) [17]. The availability of a series of homogeneous tables describing sectoral trade flows within and between countries was a significant factor behind the revitalisation of the global value chains and defragmentation studies [2, 50, 51, 79], as well as for the analysis of the global environmental footprints [3, 59].

Still, those global databases based on official national accounting data (survey) are only available at the country level, and researchers and statistical offices that want to use a multi-regional Input-Output model at a subnational level need to estimate such interregional

[2] these are non-emergency urban requests or complaints reported to the 311 phone number

flows between sectors and regions within a country. Several non-survey methods were developed with that aim, among them the ones based on location quotients, the cross-hauling adjusted regionalisaation method (CHARM), and entropy methods. They all rely on structural macroeconomics identities in order to be consistent with the total volumes coming from the known regional figures and with the sector-by-sector framework. Examples of this are the works by Sargento et al. [66], Többen and Kronenberg [81] or Boero et al. [7], among many others.

More related to the focus of this paper, Hellmanzik and Schmitz [25] and Hellmanzik and Schmitz [26] explored the role of 'virtual' proximity in explaining the trade in services between countries and their international financial integration. Both papers used data from Chung [12], who utilised the universe of the Yahoo indexed websites from 2003 and 2009: 33.8 billion websites from 273 different country top-level domains (ccTLD). They mined these data and identified 9.3 billion hyperlinks between these websites. The aggregation of these bilateral hyperlinks at country level was termed as virtual proximity. Following Kimura and Lee [37], Hellmanzik and Schmitz [25] and Hellmanzik and Schmitz [26] estimated gravity models to test whether international trade is associated with the volume of hyperlinks between countries. Their results indicated that indeed, the aggregated volume of hyperlinks between countries is

a significant determinant of services trade and its effect is particularly large for finance services, but also for communications, insurance, IT and audio-visual services. Government and construction trade services, on the other hand, appear to be less associated with virtual proximity. In any case, their findings illustrate how virtual proximity may reduce the negative effects of distance, providing a possible explanation for the decline in the distance effect on international services trade found by Head et al. [24].

Web data have been utilised before in order to study spatial relationships. Recently, Meijers and Peris [53] proposed the 'toponym co-occurrence approach' to study intercity relationships. Their study is based on retrieving relevant information from text corpora by considering when places (i.e. toponyms) are mentioned together in the same website. Then, they employ machine learning techniques to understand the context within which these place toponyms co-occurred and cluster these relationships. Their results reflect the spatial interdependencies within the Dutch settlement system and illustrate the utility of web data to capture such spatial relationships and complement existing relational data sources or substitute the lack of such data. In a similar manner, Devriendt et al. [16] and Janc [33] had employed the Google search engine to create counts of webpages which mentioned pairs of cities in order to build urban connectivity measures.

Other researchers focused on more handpicked subsets of the web. Keßler [36] and Salvini and Fabrikant [65] employed the Wikipedia as their means to study spatial relationships. While the former used the hyperlinks between German Wikipedia webpages to represent the hierarchy of urban centres in Germany, the latter utilised the English Wikipedia to build a graph of world cities. Lin et al. [47] used hyperlinks from US-based webblogs to analyse the spatial reflections of the blogsphere. In the same vein, Jones et al. [35] used hyperlinks between webblogs focused on the New York City theater scene to investigate the existence and role of a 'virtual buzz'. A number of studies utilised hyperlinks between and to administrative websites to study spatial relationships and structure [27, 28, 32].

Studying university websites and their hyperlinks is a popular application of *webometrics*, or, in other words, "the quantitative study of [w]eb-related phenomena" Thelwall et al. [74]. Early work from Thelwall [72, 73] analysed the distribution of hyperlinks between university websites and the role geography plays in how universities establish hyperlinks. Ortega and Aguillo [2008a; 2008b; 2009] broaden the scale of analysis focusing on universities from Europe and around the world. More recently, Hale et al. [23] using the same data employed for this paper analysed the graph of hyperlinks between UK university websites. Reflecting earlier results, their analysis highlighted the role of distance in establishing hyperlinks contrary to league table rankings, which do not seem to drive such linkages.

The association of physical distance with the distribution of hyperlinks was the focus of the earliest, to our knowledge, application of webometrics on studying spatial relationships. Using a limited sample of websites, Halavais [22] indicated that hyperlinks tend to follow national borders and gravitate towards the US.

The use of web data has also been employed to answer business related research questions. Vaughan et al. [85] studied hyperlinks to business websites and found that most such links reflect business motivations and contain useful business information. Nevertheless, they observed scarcity of links to competitors. Other studies found significant correlations between the number of incoming links and business performance [84, 86]. More recently, Krüger et al. [40] used hyperlinks between business websites in Germany to test the role of different proximity frameworks, which were operationalised using hyperlinks data, on the innovative behaviour of these firms. Their results indicated that innovative businesses share more hyperlinks with other business, which also tend to be innovative. Moreover, innovative businesses are being located in dense urban areas and share hyperlinks with websites from remote businesses.

In summary, the scarce of bilateral regional trade data is a well-established problem in the relevant literature. Research has done some first steps towards employing the wealth of web data in order to capture country-to-country trade relationships. Such approaches capitalised the tradition of webometrics research, which has also been focusing on illustrating spatial relationships. Building upon such studies, this paper aims to address the lack of regional trade data problem by employing open and underutilised data from web archives. Importantly, we do this within a modern ML framework, which is described in the next section.

3 METHODOLOGICAL FRAMEWORK

The main method employed here to predict inter-regional trade flows is Random Forests (RF). This is a widely used ML technique both for regression and classification problems [6]. It was firstly introduced by Breiman [8] and since then its popularity increases making it a standard tool for data science problems. The benefits of RF include its capacity to handle skewed distributions and outliers and to effectively model non-linear relationships between the dependent and independent variables, the small number of hyper-parameters that need to be tuned, the low sensitivity towards the values of these parameters as well as the relatively short training time [9, 45, 89]. Moreover, predictions based on RF are more accurate than those based on single regression trees, can illustrate the predictor importance, are fast and easy to implement [6, 8, 62, 71]. Current economic thinking advocates towards the use of RF as they tend to outperform ordinary least squares in out-of-sample predictions even when using moderate size training datasets and limited number of predictors [4, 54].

RF is a tree-based ensemble learning method [8]. It starts by creating random samples of the training data, which are then used to grow an equivalent number of regression trees to predict the dependent variable. This variable can be either a continuous one for regression problems or a categorical one for classification problems. Both observations and predictors are randomly sampled for the individual trees. Importantly, no pruning is applied for the trees to fully grow. Instead, these decision trees are trained in parallel on their own sample of the training data created with bootstrapping. An essential attribute of RF is their potential *not* to overfit. The latter refers to the capacity of a model to explain very well a specific dataset, but not being able to generalise the learned patterns to an unseen test dataset. Even though each tree on its own can overfit, the forest – i.e. the ensemble of trees – does not suffer from overfitting because the individual tree errors are averaged to produce the error of the overall forest leading also to decreased variance between trees [42]. To make a prediction for regression problems, RF average the predictions of all decision trees. So, the trade flows *IO* between regions i and j can be represented as:

$$f(X_i, X_j, Z_{ij}|\hat{\theta}|) = \frac{1}{B}\sum_{b=1}^{B}\hat{T_b} \tag{1}$$

where, X_i, X_j, and Z_{ij} are the origin i, destination j and origin-destination pair ij predictors of regional trade respectively, $\hat{\theta}$ is a vector of the estimated hyperparameters, B is the total number of trees the RF grew, which is also one of the estimated hyperparameters and $\hat{T_b}$ represents the prediction of each independent tree [89].

To estimate RF models we employ the widely used `caret` package for `R` [41] and we build the following *rolling forecasting* workflow: *(1)* train RF models on data from years t and $t + 1$ from the study period 2000-2010 to increase the size of the training dataset; *(2)* evaluate their predictive capability using cross validation (CV); *(3)* apply the estimated RF models from step *(1)* on unseen data from the following years $(t + 2)$ to predict trade flows for that year and evaluate their predictive capability of such unseen data.

We opted against pooling the data to maintain their temporal structure both for methodological and conceptual reasons. Random sampling and pooling may lead to utilise future values of hyperlink flows to predict past regional trade flows, which might be counter-intuitive.Instead, we opted towards building biyearly RF modesl with 10-fold CV to assess their predictive capability. To do so, we split the biyearly subsets of the data in 10 random samples, trained a RF on the nine and tested its predictive capacity on the holdout one.This process is repeated ten times in order for all 10 samples to act as holdout ones. Then, the predictive performance of the RF is equal to the mean performance of the 10 models. This workflow enables us to make out-of-sample predictions and test our models and research framework in previously unseen data. Successful predictions within this framework will illustrate the digital traces that regional trade leaves behind and how useful such web data are in predicting trade flow that we have little information about.

To assess the predictive capability of the models, we utilise three broadly used metrics: mean absolute error (MAE), root mean square error (RMSE) and the coefficient of determination (R squared):

$$R^2 = 1 - \frac{\sum_k (y_k - \hat{y_k})^2}{\sum_k (y_k - \overline{y_k})^2} \tag{2}$$

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|\hat{y_k} - y_k| \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N}(\hat{y_k} - y_k)^2}{N}} \tag{4}$$

y_k is the k^{th} observation of the dataset, which consists of N observation in total. $\hat{y_k}$ is the k_{th} predicted value for the dependent variable and $\overline{y_k}$ is the average value of y. The last two metrics are expressed in the same units as the dependent variable – $\$100,000s$ – while the first one is the coefficient of determination between the observed and the predicted values of regional trade flows. Regarding *MAE*, it is the absolute difference between the observed and the predicted trade flows. While *MAE* does not penalise for large errors, *RSME* does so as it is proportional to the squared difference between the observed and the predicted trade flows. This means that larger errors carry more weight for *RMSE* [61].

RF allow to estimate the predictor importance. Using the built-in bootstrapping procedure, MSE is recorded for each tree when including all the predictors and then again by excluding one by one all the predictors. The derived decrease in the MSE created by removing each predictor signifies its importance in the model [8][3].

We then move to test the predictive capacity of RF trained on year t and $t+1$ on unseen data from year $t+2$. This yearly data split equites to a *firewall principle* [54], which prevent data leakage between the training and the test dataset. Being able to accurately predict unseen regional trade flows provides further statistical evidence for the validity of for our proposed research strategy. Moreover, it advocates towards the utility of our research strategy and its applicability in predicting regional trade flows.

Our research framework enables us to disaggregate the models based on economic sectors and, therefore, we are able to further assess their sensitivity. In addition, we utilise different subsets of the web data as another robustness checks. The details of these data are exaplined in the next section.

4 DATA

The hyperlinks data have been derived from the JISC UK Web Domain Datasetˆ, [https://data.webarchive.org.uk/opendata/ukwa.ds.2/] [34], which contains all the archived .uk webpages contained in the Internet Archive. These data have been utilised in other

[3] see also the `randomForest` `R` package, which is based on Breiman's [2001] original implementation

studies which explored the role of online content of local interest in attracting individuals online [83] and on the long term regional productivity effects of the early adoption of web technologies [82].

In the case of the information on interregional trade, we obtain the flows of imports and exports between the UK NUTS2 regions from the EUREGIO database (Thissen et al. [78]), which are the most detailed data currently available about the economic and trade structure of the UK and EU regions. The EUREGIO uses the World Input-Output Database (WIOD) (Timmer et al. [79]) as a starting point and adds regional detail for EU Member States as of 2010. The EUREGIO is available for the years 2000 to 2010, and it contains information for 256 European NUTS2 regions and 14 sectors in each region (Ijtsma et al. [29]).

Regional trade in the EUREGIO database is taken from the PBL Netherlands Environmental Assessment Agency regional trade data for the year 2000 as a prior to the estimations for the whole series 2000-2010 (Thissen et al. [77] and Thissen et al. [76]). The PBL/RT dataset was constructed by merging data from several sources: national accounts of the selected countries; international trade data on goods from Feenstra et al. [19] and on services from Eurostat; macroeconomic regional data from Cambridge Econometrics and Eurostat's regional accounts; information on freight transport among European regions for approximating the network of trade in goods; and first and business class airline tickets information for approximating the network of trade in services. Therefore, in the EUREGIO database no spatial structure has been imposed on the data, which means that no specific model was used to estimate trade flows and patterns. The procedure used allocates the trade over the regions depending on the amounts produced and consumed in every region. The estimation approach ensures the final consistency of the regional tables with the national tables (Thissen et al. [78]; Ivanova et al. [31]).

This database has been used recently in studies estimating the impacts of different economic shocks. Los et al. [49], paradoxically found that those regions that voted in favour of leaving the EU in the 2016 Brexit referendum were the ones with a higher share of local economic activity dependent on the trade with the EU, and therefore the ones that would suffer more the negative economic consequences of a rupture scenario. Similarly, Chen et al. [10] used these data to estimate the regional exposure to Brexit for the whole European Union. Kitsos et al. [38] employed these data to examine the role of local industrial embeddedness on economic resilience for the UK regions. Wilting et al. [88], among others, estimated the subnational greenhouse gas and land-based biodiversity footprints in the EU regions using this database.

5 RESULTS

The first step of the analysis involves training our RF models using data from years t and $t + 1$. Figure 1 presents the accuracy metrics for the training data set that were obtained through the 10-fold CV. The results clearly indicate that our models achieve high in-sample accuracy across all three prediction accuracy metrics employed here. There is some variation between years, but still the results appear to be very promising.

To avoid overfitting and test the predictive capacity of our models on unseen data the models estimated using data from years t and $t + 1$ are aplied on years $t + 2$. The yearly accuracy metrics are presented in Table 1 and the predicted versus the observed flows of interregional trade are plotted in Figure 2. Indeed, our models are able to make highly accurate out of sample predictions for interregional trade in the UK. The R-squared only drops below 0.9 in 2005 and 2010 (0.89 and 0.63), while it exceeds 0.95 in 2002 and 2004. As Figure 2 indicates the highest errors are observed for the regional pairs with the two highest flows of interregional trade every year and our models under- and overestimate their flows. These outliers are always the intra-regional flows within Inner and Outer London regions (UKI1 and UKI2). With the exception of these extreme values though (trade flows above $\$50\ billions$) our model perform remarkably well in out of sample predictions. Comparing to other attempts in the literature to predict trade flows.

Table 1: Accuracy metrics in unseen data from t + 2

year	RMSE	Rsquared	MAE
2002	937.93	0.96	159.87
2003	1360.28	0.94	244.75
2004	1014.83	0.95	179.15
2005	1790.07	0.89	304.86
2006	1706.73	0.92	309.16
2007	1920.11	0.91	210.23
2008	1558.92	0.92	233.35
2009	1353.12	0.93	202.7
2010	3170.16	0.63	303.68

Our research framework enables us to disaggregate our results in terms of economic sectors. So, we repeat the same modelling procedure for each sector separately and Figure 3 reports the R-squared values for the predictions of the unseen $t + 2$ interregional trade flows for each sector. As expected, our model perform worst in specific sectors. The most obvious example is hospitality, the R-squared for which drops below 0.25 between the predicted and observed values for 2004 and 2010. Through the study period though it does not exceed 0.75. A similar trend – although not as dramatic – can be observed for real estate. With the exception of 2010, the predictive capacity of our models is high for all the sectors as R-squared is consistently above .075. Our models perform exceptionally good in predicting unseen interregional trade flows regarding manufacturing and construction as well as requirement.

To further assess the role of our main variable of interest – the volume of hyperlinks between regions – in predicting interregional trade flows we estimate the first set of models for the total trade flows using alternative specifications. Firstly, we exclude the distance or the hyperlinks features. The accuracy metrics for the out of sample predictions for unseen trade flow data from years $t + 2$ are presented in Figure **??**, which also includes the metrics for the base models presented in Table 1 for direct comparison. The main message from Figure **??** is that distance plays the most important role in predicting interregional trade flows. All three metrics are worst when the distance is excluded. This is not surprising as the role of distance in predicting trade and other types of spatial interactions has been extensively highlighted in the literature discussed in Sections 1 and 2. What is interesting though is that the gap in terms of the prediction accuracy between the models with and without distance decreases over time. This illustrates that over time,

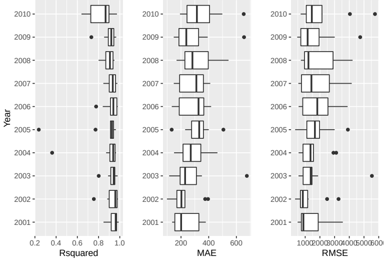

Figure 1: Accuracy metrics

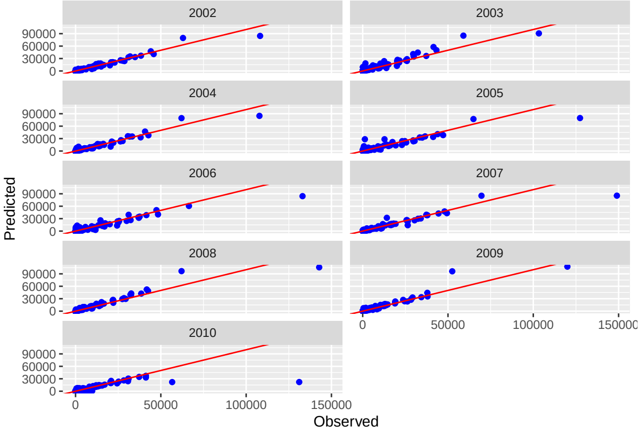

Figure 2: Predicted vs. observed interregional trade by year

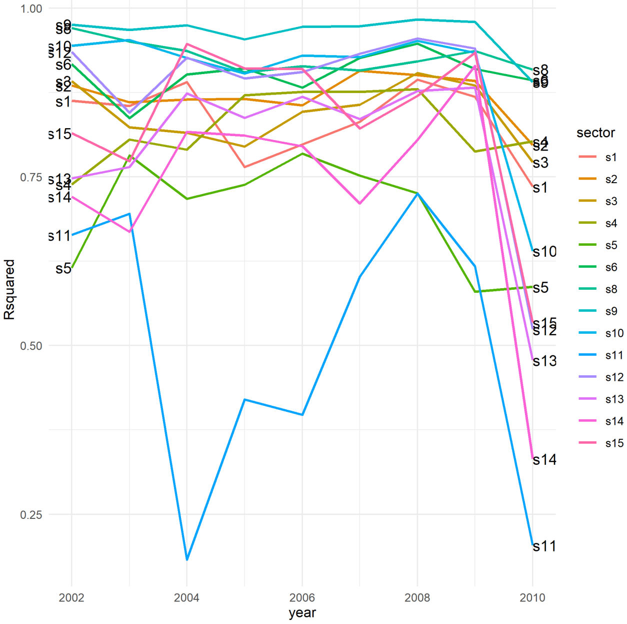

Figure 3: R-squared for t + 2 out of sample predictions per sector.

Notes: s1: Agriculture, s2: Mining, s3: Food, s4: Textiles, s5: Chemicals, s6: Equipment, s8: Manufacturing; s9: Construction, s10: Distribution, s11: Hospitality, s12: Transport, s13: Financial, s14: Real Estate, s15: Non-Market Services.

as the adoption rate of web technologies increased, interregional trade flows started leaving more 'digital breadcrumbs' behind and, therefore, are better reflected in the volumes of interregional hyperlinks [63]. Nevertheless, the predictive capacity of distance remains unchallenged at large as the green lines in Figure **??** indicate.

To further assess the robustness of our results, we repeat our workflow for a different sample of websites. Instead of including only the websites with a unique postcode, we add in our sample websites with up to 11 unique postcodes. As discussed in Section 4, this enhanced sample of websites containing multiple postcodes within the web text represents commercial websites with multiple locations. Given that we are not able to distinguish the role of these different locations we expect that using this sample for training and testing our models will lead to more noise. Nevertheless, the predictive capacity of our models remains almost unchanged according to the out of sample prediction metrics, which are reported in Table 2 and the predicted versus the observed interregional trade flows, which are plotted in Figure **??**. Indeed, R-squared drops below 0.90 for only two years (2009 and 2010). Again, the largest prediction errors can be easily linked to the regional pairs with the highest volume of regional trade – that is the London intra-regional trade.

Table 2: Accuracy metrics in unseen data with multiple postcodes from t + 2

year	RMSE	Rsquared	MAE
2002	1181.91	0.94	244.27
2003	1428.99	0.93	282.77
2004	1011.14	0.95	173.31
2005	1414.77	0.94	232.25
2006	1433.92	0.94	208.32
2007	1894.59	0.91	227.77
2008	1206.3	0.95	249.66
2009	2008.83	0.81	238.38
2010	2500.1	0.78	298.27

6 CONCLUSIONS

In summary, it is very well established in the literature that interregional trade is difficult to capture. The current state of the art of empirical studies focusing on modelling international trade tend to be explanatory in their nature and most attempts to produce disaggregated interregional trade flows are mostly based on pure distance decay measures. This paper proposes the use of innovative and openly accessible web data within a state-of-the art ML framework to predict such trade flows. The predictive capacity of our model is very good and it is indicative of the *digital* paper trail that trade leaves behind. Our framework could be utilised in framework to produce such interregional data, which otherwise would be very difficult to capture, and also produce these data at much more disaggregated scales than before.

7 REFERENCES

REFERENCES

[1] James E Anderson and Eric Van Wincoop. 2003. Gravity with gravitas: A solution to the border puzzle. *American economic review* 93, 1 (2003), 170–192.

[2] Pol Antras and Davin Chor. 2018. *On the measurement of upstreamness and downstreamness in global value chains.* Technical Report. National Bureau of Economic Research.

[3] Iñaki Arto, José M Rueda-Cantuche, and Glen P Peters. 2014. Comparing the GTAP-MRIO and WIOD databases for carbon footprint analysis. *Economic Systems Research* 26, 3 (2014), 327–353.

[4] Susan Athey and Guido W Imbens. 2019. Machine learning methods that economists should know about. *Annual Review of Economics* 11 (2019), 685–725.

[5] Fabrizio Barca. 2009. An Agenda for a Reformed Cohesion Policy-Independent Report. *European Commission, Brussels* (2009).

[6] GÃŠrard Biau. 2012. Analysis of a random forests model. *Journal of Machine Learning Research* 13, Apr (2012), 1063–1095.

[7] Riccardo Boero, Brian K Edwards, and Michael K Rivera. 2018. Regional input–output tables and trade flows: an integrated and interregional non-survey approach. *Regional Studies* 52, 2 (2018), 225–238.

[8] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[9] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. 2008. An Empirical Evaluation of Supervised Learning in High Dimensions. In *Proceedings of the 25th International Conference on Machine Learning* (Helsinki, Finland) *(ICML '08)*. Association for Computing Machinery, New York, NY, USA, 96–103. https://doi.org/10.1145/1390156.1390169

[10] Wen Chen, Bart Los, Philip McCann, Raquel Ortega-Argilés, Mark Thissen, and Frank van Oort. 2018. The continental divide? Economic exposure to Brexit in regions and countries on both sides of The Channel. *Papers in Regional Science* 97, 1 (2018), 25–54.

[11] Yongwan Chun, Hyun Kim, and Changjoo Kim. 2012. Modeling interregional commodity flows with incorporating network autocorrelation in spatial interaction models: An application of the US interstate commodity flows. *Computers, Environment and Urban Systems* 36, 6 (2012), 583–591.

[12] Joo Chung. 2011. *The geography of global internet hyperlink networks and cultural content analysis.* Ph.D. Dissertation. Dissertation, University at Buffalo.

[13] Kevin Credit. [n.d.]. Spatial Models or Random Forest? Evaluating the Use of Spatially Explicit Machine Learning Methods to Predict Employment Density around New Transit Stations in Los Angeles. *Geographical Analysis* ([n. d.]).

[14] Felipa de Mello-Sampayo. 2017. Competing-destinations gravity model applied to trade in intermediate goods. *Applied Economics Letters* 24, 19 (2017), 1378–1384.

[15] Felipa De Mello-Sampayo. 2017. Testing competing destinations gravity models–evidence from BRIC International. *The Journal of International Trade & Economic Development* 26, 3 (2017), 277–294.

[16] Lomme Devriendt, Ben Derudder, and Frank Witlox. 2008. Cyberplace and cyberspace: two approaches to analyzing digital intercity linkages. *Journal of Urban Technology* 15, 2 (2008), 5–32.

[17] Erik Dietzenbacher, Bart Los, Robert Stehrer, Marcel Timmer, and Gaaitzen De Vries. 2013. The construction of world input–output tables in the WIOD project. *Economic Systems Research* 25, 1 (2013), 71–98.

[18] Peter Egger. 2002. An econometric view on the estimation of gravity models and the calculation of trade potentials. *World Economy* 25, 2 (2002), 297–312.

[19] Robert C Feenstra, Robert E Lipsey, Haiyan Deng, Alyson C Ma, and Hengyong Mo. 2005. *World trade flows: 1962-2000.* Technical Report. National Bureau of Economic Research.

[20] Estrella Gómez-Herrera. 2013. Comparing alternative methods to estimate gravity models of bilateral trade. *Empirical economics* 44, 3 (2013), 1087–1111.

[21] Raf Guns and Ronald Rousseau. 2014. Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics* 101, 2 (2014), 1461–1473.

[22] Alexander Halavais. 2000. National borders on the world wide web. *New Media & Society* 2, 1 (2000), 7–28.

[23] Scott A Hale, Taha Yasseri, Josh Cowls, Eric T Meyer, Ralph Schroeder, and Helen Margetts. 2014. Mapping the UK webspace: Fifteen years of british universities on the web. In *Proceedings of the 2014 ACM conference on Web science.* 62–70.

[24] Keith Head, Thierry Mayer, and John Ries. 2009. How remote is the offshoring threat? *European Economic Review* 53, 4 (2009), 429–444.

[25] Christiane Hellmanzik and Martin Schmitz. 2016. Gravity and international services trade: the impact of virtual proximity. *Eur. Econ. Rev* 77 (2016), 82–101.

[26] Christiane Hellmanzik and Martin Schmitz. 2017. Taking gravity online: The role of virtual proximity in international finance. *Journal of International Money and Finance* 77 (2017), 164–179.

[27] Kim Holmberg. 2010. Co-inlinking to a municipal Web space: a webometric and content analysis. *Scientometrics* 83, 3 (2010), 851–862.

[28] Kim Holmberg and Mike Thelwall. 2009. Local government web sites in Finland: A geographic and webometric analysis. *Scientometrics* 79, 1 (2009), 157–169.

[29] Pieter Ijtsma, Bart Los, et al. 2020. *UK Regions in Global Value Chains.* Technical Report. Economic Statistics Centre of Excellence (ESCoE).

[30] Walter Isard. 1956. Location and space-economy. (1956).

[31] Olga Ivanova, d'Artis Kancs, and Mark Thissen. 2019. *Regional Trade Flows and Input Output Data for Europe.* Technical Report. EERI Research Paper Series.

[32] Krzysztof Janc. 2015. Geography of hyperlinks—Spatial dimensions of local government websites. *European Planning Studies* 23, 5 (2015), 1019–1037.

[33] Krzysztof Janc. 2015. Visibility and connections among cities in digital space. *Journal of Urban Technology* 22, 4 (2015), 3–21.

[34] JISC and the Internet Archive. [n.d.]. JISC UK Web Domain Dataset (1996-2013). *The British Library* ([n. d.]). https://doi.org/10.5259/ukwa.ds.2/1

[35] Brant W Jones, Ben Spigel, and Edward J Malecki. 2010. Blog links as pipelines to buzz elsewhere: the case of New York theater blogs. *Environment and Planning B: Planning and Design* 37, 1 (2010), 99–111.

[36] Carsten Keßler. 2017. Extracting central places from the link structure in Wikipedia. *Transactions in GIS* 21, 3 (2017), 488–502.

[37] Fukunari Kimura and Hyun-Hoon Lee. 2006. The gravity equation in international trade in services. *Review of world economics* 142, 1 (2006), 92–121.

[38] Anastasios Kitsos, André Carrascal-Incera, and Raquel Ortega-Argilés. 2019. The role of embeddedness on regional economic resilience: Evidence from the UK. *Sustainability* 11, 14 (2019), 3800.

[39] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–95.

[40] Miriam Krüger, Jan Kinne, David Lenz, and Bernd Resch. 2020. The digital layer: How innovative firms relate on the web. *ZEW-Centre for European Economic Research Discussion Paper* 20-003 (2020).

[41] Max Kuhn et al. 2008. Building predictive models in R using the caret package. *Journal of statistical software* 28, 5 (2008), 1–26.

[42] Mark Last, Oded Maimon, and Einat Minkov. 2002. Improving stability of decision trees. *International Journal of Pattern Recognition and Artificial Intelligence* 16, 02 (2002), 145–159.

[43] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Social science. Computational social science. *Science (New York, NY)* 323, 5915 (2009), 721–723.

[44] E. Leamer and R. Stern. 1971. Quantitative International Economics. *Journal of International Economics* 1 (1971), 359–361.

[45] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[46] Marcio Salles Melo Lima and Dursun Delen. 2020. Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly* 37, 1 (2020), 101407.

[47] Jia Lin, Alexander Halavais, and Bin Zhang. 2007. The blog network in America: blogs as indicators of relationships among US cities. *Connections* 27, 2 (2007), 15–23.

[48] Hans Linnemann. 1966. *An econometric study of international trade flows.* Number 42. North-Holland Pub. Co.

[49] Bart Los, Philip McCann, John Springford, and Mark Thissen. 2017. The mismatch between local voting and the local economic consequences of Brexit. *Regional Studies* 51, 5 (2017), 786–799.

[50] Bart Los, Marcel P Timmer, and Gaaitzen J de Vries. 2015. How global are global value chains? A new approach to measure international fragmentation. *Journal of regional science* 55, 1 (2015), 66–92.

[51] Bart Los, Marcel P Timmer, and Gaaitzen J de Vries. 2016. Tracing value-added and double counting in gross exports: comment. *American Economic Review* 106, 7 (2016), 1958–66.

[52] Philip McCann and Raquel Ortega-Argilés. 2015. Smart specialization, regional growth and applications to European Union cohesion policy. *Regional studies* 49, 8 (2015), 1291–1302.

[53] Evert Meijers and Antoine Peris. 2019. Using toponym co-occurrences to measure relationships between places: review, application and evaluation. *International Journal of Urban Sciences* 23, 2 (2019), 246–268.

[54] Sendhil Mullainathan and Jann Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31, 2 (2017), 87–106.

[55] José Luis Ortega and Isidro F Aguillo. 2008. Linking patterns in European Union countries: geographical maps of the European academic web space. *Journal of Information Science* 34, 5 (2008), 705–714.

[56] José Luis Ortega and Isidro F Aguillo. 2008. Visualization of the Nordic academic web: Link analysis using social network tools. *Information Processing & Management* 44, 4 (2008), 1624–1633.

[57] Jose Luis Ortega and Isidro F Aguillo. 2009. Mapping world-class universities on the web. *Information Processing & Management* 45, 2 (2009), 272–279.

[58] Taylor M Oshan. 2020. The spatial structure debate in spatial interaction modeling: 50 years on. *Progress in Human Geography* (2020), 0309132520968134.

[59] Anne Owen, Richard Wood, John Barrett, and Andrew Evans. 2016. Explaining value chain differences in MRIO databases through structural path decomposition. *Economic Systems Research* 28, 2 (2016), 243–272.

[60] James Paul Lesage and Wolfgang Polasek. 2008. Incorporating transportation network structure in spatial econometric models of commodity flows. *Spatial Economic Analysis* 3, 2 (2008), 225–245.

[61] Robert Gilmore Pontius, Olufunmilayo Thontteh, and Hao Chen. 2008. Components of information for multiple resolution comparison between maps that share a real variable. *Environmental and Ecological Statistics* 15, 2 (2008), 111–142.

[62] Nastaran Pourebrahim, Selima Sultana, Amirreza Niakanlahiji, and Jean-Claude Thill. 2019. Trip distribution modeling with Twitter data. *Computers, Environment*

and Urban Systems 77 (2019), 101354.

[63] Chirag Rabari and Michael Storper. 2014. The digital skin of cities: urban theory and research in the age of the sensored and metered city, ubiquitous computing and big data. *Cambridge Journal of Regions, Economy and Society* 8, 1 (10 2014), 27–42. https://doi.org/10.1093/cjres/rsu021

[64] Yi Ren, Tong Xia, Yong Li, and Xiang Chen. 2019. Predicting socio-economic levels of urban regions via offline and online indicators. *PloS one* 14, 7 (2019).

[65] Marco M Salvini and Sara I Fabrikant. 2016. Spatialization of user-generated content to uncover the multirelational world city network. *Environment and Planning B: Planning and Design* 43, 1 (2016), 228–248.

[66] Ana LM Sargento, Pedro Nogueira Ramos, and Geoffrey JD Hewings. 2012. Interregional trade flow estimation through non-survey models: An empirical assessment. *Economic Systems Research* 24, 2 (2012), 173–193.

[67] Ma Ángeles Serrano and Marián Boguñá. 2003. Topology of the world trade web. *Phys. Rev. E* 68 (Jul 2003), 015101. Issue 1. https://doi.org/10.1103/PhysRevE.68.015101

[68] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. 2012. A universal model for mobility and migration patterns. *Nature* 484, 7392 (2012), 96–100.

[69] Alex Singleton and Daniel Arribas-Bel. 2021. Geographic data science. *Geographical Analysis* 53, 1 (2021), 61–75.

[70] Parmanand Sinha, Andrea E Gaughan, Forrest R Stevens, Jeremiah J Nieves, Alessandro Sorichetta, and Andrew J Tatem. 2019. Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. *Computers, Environment and Urban Systems* 75 (2019), 132–145.

[71] Sarina Sulaiman, Siti Mariyam Shamsuddin, Ajith Abraham, and Shahida Sulaiman. 2011. Intelligent web caching using machine learning methods. *Neural Network World* 21, 5 (2011), 429.

[72] Mike Thelwall. 2002. Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation* (2002).

[73] Mike Thelwall. 2002. The top 100 linked-to pages on UK university web sites: high inlink counts are not usually associated with quality scholarly content. *Journal of information science* 28, 6 (2002), 483–491.

[74] Mike Thelwall, Liwen Vaughan, and Lennart Björneborn. 2005. Webometrics. *Annual Review of Information Science and Technology* 39, 1 (2005), 81–135. https://doi.org/10.1002/aris.1440390110 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440390110

[75] Mark Thissen, Thomas de Graaff, and Frank van Oort. 2016. Competitive network positions in trade and structural economic growth: A geographically weighted regression analysis for European regions. *Papers in Regional Science* 95, 1 (2016), 159–180.

[76] M Thissen, D Diodato, and F Van Oort. 2013. European regional trade flows: An update for 2000–2010. *PBL Netherlands Environmental Assessment Agency, The Hague* (2013).

[77] M Thissen, D Diodato, and F Van Oort. 2013. Integrated regional Europe: European regional trade flows in 2000. *PBL Netherlands Environmental Assessment Agency, The Hague* (2013).

[78] Mark Thissen, Maureen Lankhuizen, Frank van Oort, Bart Los, and Dario Diodato. 2018. EUREGIO: The construction of a global IO DATABASE with regional detail for Europe for 2000–2010. (2018).

[79] Marcel P Timmer, Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J De Vries. 2015. An illustrated user guide to the world input–output database: the case of global automotive production. *Review of International Economics* 23, 3 (2015), 575–605.

[80] Jan Tinbergen. 1962. Shaping the World Economy The Twentieth Century Fund. *New York* 330 (1962).

[81] Johannes Többen and Tobias Heinrich Kronenberg. 2015. Construction of multiregional input–output tables using the CHARM method. *Economic systems research* 27, 4 (2015), 487–507.

[82] Emmanouil Tranos, Tasos Kitsos, and Raquel Ortega-Argilés. 2021. Digital economy in the UK: Regional productivity effects of early adoption. *Regional Studies* in press (2021).

[83] Emmanouil Tranos and Christoph Stich. 2020. Individual internet usage and the availability of online content of local interest: A multilevel approach. *Computers, Environment and Urban Systems* 79 (2020), 101371.

[84] Liwen Vaughan. 2004. Exploring website features for business information. *Scientometrics* 61, 3 (2004), 467–477.

[85] Liwen Vaughan, Yijun Gao, and Margaret Kipp. 2006. Why are hyperlinks to business Websites created? A content analysis. *Scientometrics* 67, 2 (2006), 291–300.

[86] Liwen Vaughan and Guozhu Wu. 2004. Links to commercial websites as a source of business information. *Scientometrics* 60, 3 (2004), 487–496.

[87] Lingjing Wang, Cheng Qian, Philipp Kats, Constantine Kontokosta, and Stanislav Sobolevsky. 2017. Structure of 311 service requests as a signature of urban location. *PloS one* 12, 10 (2017).

[88] Harry C Wilting, Aafke M Schipper, Olga Ivanova, Diana Ivanova, and Mark AJ Huijbregts. 2020. Subnational greenhouse gas and land-based biodiversity footprints in the European Union. *Journal of Industrial Ecology* (2020).

[89] Xiang Yan, Xinyu Liu, and Xilei Zhao. 2020. Using machine learning for direct demand modeling of ridesourcing services in Chicago. *Journal of Transport Geography* 83 (2020), 102661.