Using the web to predict regional trade flows: data extraction, modelling, and validation

Emmanouil Tranos[a,b,†], Andre Carrascal Incera[c], George Willis[d]

[a]University of Bristol, UK; [b]The Alan Turing Institute, UK; [c]University of Oviedo, Spain; [d]University of Birmingham, UK; [†]Corresponding author

ARTICLE HISTORY
Compiled June 13, 2022

ABSTRACT
Despite the importance of interregional trade for building effective regional economic policies, there is very little hard data to illustrate such interdependencies. We propose here a novel research framework to predict interregional trade flows by utilising freely available web data and machine learning algorithms. Specifically, we extract hyperlinks between archived websites in the UK and we aggregate these data to create an interregional network of hyperlinks between geolocated and commercial webpages over time. We also use some existing interregional trade data to train our models using random forests and then make out-of-sample predictions of interregional trade flows using a rolling-forecasting framework. Our models illustrative great predictive capability with R^2 greater than 0.9. We are also able to disaggregate our predictions in terms of industrial sectors, but also at a sub-regional level, for which trade data are not available. In total, our models provide a proof of concept that the digital traces left behind by physical trade can help us capture such economic activities at a more granular level and, consequently, inform regional policies.

KEYWORDS
interregional trade; web archives; web data; machine learning; prediction; random forest

1. Introduction

Bilateral trade is a complex phenomenon per se (Serrano and Boguñá 2003), but its complexity increases when it is approached from a spatially disaggregated perspective. Regions[1] behave differently 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 (Isard 1951; Miller and Blair 2009). Therefore, they face intense external dependencies (Matter 2009). 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 (Fingleton, Garretsen, and Martin 2012). 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

CONTACT Emmanouil Tranos. Email: e.tranos@bristol.ac.uk, Andre Carrascal Incera. Email: carrascalandre@uniovi.es, George Willis. Email: GCW519@student.bham.ac.uk

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

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 (Guan et al. 2020; David, Dorn, and Hanson 2013). 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 (Thissen, de Graaff, and van Oort 2016). This is why, among other factors, regions had significantly divergent experiences in avoiding or overcoming economic shocks (Kitsos, Carrascal-Incera, and Ortega-Argilés 2019).

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 (Barca 2009). Interregional relations and modern supply chains are central in a systemic way of thinking about regional innovation and growth strategies (Thissen, Diodato, and Van Oort 2013b) such as the smart specialisation policy initiatives (McCann and Ortega-Argilés 2015). Equally, not having a clear picture of regional trade dependencies may impede our capacity to design effective regional economic policies. Our paper provides tools to map such regional trade dependencies.

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, Diodato, and Van Oort (2013b), who followed the parameter-free Simini et al. (2012) 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. Normally, these type of data are only available at the national level and released infrequently (usually every 5 years) by statistical institutes, because they are based in expensive and time-consuming industrial surveys (Boero, Edwards, and Rivera 2018). This illustrates the difficulty of building a database about interregional trade.

Our paper contributes to this line of inquiry by utilising novel web data and machine learning algorithms to make temporal 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 commercial 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 physical trade leaves behind digital breadcrumbs (Rabari and Storper 2014), which can be effectively utilised to capture interregional trade flows, which are both important for regional policies and also very difficult to observe.

Our proposed research framework not only allows for accurate prediction of interregional trade flows, but also for disaggregating such flows at more granular spatial units representing local authorities. Hence, it has the potential to directly support local authorities in the efforts to identify external dependencies and vulnerabilities to supply chain disruptions. Importantly, such accurately predicted interregional and granular trade flows can assist ex ante evaluations of place-based economic policies.

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 Oshan 2020b). Methodological and conceptual developments on *spatial interaction models* have been extensively employed in order to model flows of trade between regions (Chun, Kim, and Kim 2012; Paul Lesage and Polasek 2008) and countries (De Mello-Sampayo 2017; de Mello-Sampayo 2017). Following current debates within the quantitative geographical thinking (Singleton and Arribas-Bel 2021) and, more broadly, computational social sciences (Lazer et al. 2009), geographical research has been focusing more on explaining the determinants of interregional trade flows than predicting such flows.

This paper is aligned with current epistimological debates in geography (Singleton and Arribas-Bel 2021; Credit 2021) and economics (Kleinberg et al. 2015) 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 (Mullainathan and Spiess 2017; Athey and Imbens 2019). Such an approach can be particularly useful for predicting interregional trade flows given the scarcity and cost to produce such data.

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 as well as sensitivity checks. We also present an illustrative example of how our models can be used to map trade flows at an even more spatially disaggregated level. The paper ends with a conclusions section.

2. From the lack of regional trade data to the use of webdata

This section reviews different literatures, which either aim to model trade flows or employed some form of web data to capture spatial relationships given the lack of relevant data. Not having directly available data to map interregional trade hinders policy makers from understanding in detail the economic dependencies of regions and, therefore, design appropriate regional economic policies.

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 (1962), Linnemann (1966) and Leamer and Stern (1971). 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.) (Egger 2002; Anderson and Van Wincoop 2003). 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 (2013).

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 (Isard 1956), 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) (Dietzenbacher et al. 2013). 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 (Timmer et al. 2015; Los, Timmer, and de Vries 2015, 2016; Antras and Chor 2018), as well as for the analysis of the global environmental footprints (Arto, Rueda-Cantuche, and Peters 2014; Owen et al. 2016).

Still, those global databases based on official national accounting data that is business surveys 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 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 regionalisation 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, Ramos, and Hewings (2012), Többen and Kronenberg (2015) or Boero, Edwards, and Rivera (2018), among many others.

More related to the focus of this paper, Hellmanzik and Schmitz (2016) and Hellmanzik and Schmitz (2017) 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 (2011), 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 (2006), Hellmanzik and Schmitz (2016) and Hellmanzik and Schmitz (2017) 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, Mayer, and Ries (2009).

More broadly, web data have been utilised in order to study spatial relationships. Recently, Meijers and Peris (2019) 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, Derudder, and Witlox (2008) and Janc (2015b) 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 (2017) and Salvini and Fabrikant (2016) employed Wikipedia 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, Halavais, and Zhang (2007) used hyperlinks from US-based webblogs to analyse the spatial reflections of the blogsphere. In the same vein, Jones, Spigel, and Malecki (2010) 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 (Holmberg and Thelwall 2009; Holmberg 2010; Janc 2015a).

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, Vaughan, and Björneborn (2005). Early work from Thelwall (Thelwall 2002b,a) 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. (2014a) 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 (2000) 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, Gao, and Kipp (2006) studied hyperlinks to business websites and found that 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 (Vaughan 2004; Vaughan and Wu 2004). More recently, Krüger et al. (2020) 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 state-of-the-art ML framework, which is described in the next section.

3. Methodological framework

Random Forest (RF) is the main estimator we employ in order to predict inter-regional trade flows. This is a widely used ML technique both for regression and classification problems (Biau 2012). It was firstly introduced by Breiman (2001) 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 hyperparameters that need to be tuned, the low sensitivity towards the values of these parameters as well as the relatively short training time (Caruana, Karampatziakis, and Yessenalina 2008; Liaw, Wiener et al. 2002; Yan, Liu, and Zhao 2020). Moreover, predictions based on RF are more accurate than those based on single regression trees, can illustrate the predictor importance, and are fast and easy to implement (Breiman 2001; Sulaiman et al. 2011; Pourebrahim et al. 2019; Biau 2012). 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 (Mullainathan and Spiess 2017; Athey and Imbens 2019). All the above advocate towards utilising RF in this paper as we aim to do temporal out-of-sample predictions of data, which are skewed and have outliers (see for instance Figure 3).

RF have been extensively utilised in answering regression research problems. Pourebrahim et al. (2019) 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 (2020) used ML and RF to predict and explain corruption across countries. Sinha et al. (2019) 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 (2021) introduced spatially explicit RF to predict employment density in Los Angeles. Guns and Rousseau (2014) use RF to predict and recommend high-potential research collaborations, which have not yet been materialised. Ren et al. (2019) trained RF as well as other classifiers to predict socio-economic status or cities using a variety of online and mobility predictors. Wang et al. (2017) employed RF and other machine learning algorithms to predict urban socio-economic characteristic based on non-emergency urban requests or complaints reported to the 311 phone number. 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.

RF is a tree-based ensemble learning algorithm (Breiman 2001). 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 capacity *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 (Last, Maimon, and Minkov 2002). 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 (Yan, Liu, and Zhao 2020).

To estimate RF models we employ the widely used `caret` package for `R` (Kuhn et al. 2008) 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 year $(t+2)$ 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 counterintuitive. Instead, we opted towards building biyearly RF models with 10-fold CV to assess their in-sample 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 temporal out-of-sample predictions and test our models and research framework in previously unseen data. Importantly, it helps avoid overfitting, which would have occurred if the temporal structure of the data and the underpinning time-dependent data generation processes had been ignored. Such discussions can be found in the spatial interaction modelling literature (Mikkonen and Luoma 1999; Mozolin, Thill, and Usery 2000; Oshan 2020a). Successful predictions within this framework can 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: the coefficient of determination (R squared), mean absolute error (MAE) and root mean square error (RMSE):

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

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

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

y_k is the k^{th} observation of the dataset, which consists of N observations 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 $RMSE$ (Pontius, Thontteh, and Chen 2008).

RF allow the estimation of 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 (Breiman 2001)[2].

We then move to test the predictive capacity of RF trained on years t and $t+1$ on unseen data from the year $t+2$. This yearly data split equites to a *firewall principle* (Mullainathan and Spiess 2017), 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 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 and predictions in terms of both economic sectors and spatial units 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 explained in the next section.

4. Data

The hyperlinks data have been derived from the JISC UK Web Domain Dataset (JISC and the Internet Archive 2013)[3]. The latter contains all the webpages under the .uk country code top level domain (ccTLD), which have been discovered and archived by the Internet Archive (IA)[4] during the 1996-2010 period. Regarding the extent of the IA, this is not trivial to assess given that the actual size of the web is unknown. Studies from the digital humanities field claim that although it is difficult to evaluate the coverage of web archives, the IA is the most extensive and complete archive (Ainsworth et al. 2011; Holzmann, Nejdl, and Anand 2016). Using a different subset of the IA, Thelwall and Vaughan (2004) suggested that 92% of all US based commercial websites had been archived.

This openly accessible dataset contains c. 2.5 billion URLs, which point to archived .uk webpages, the HTML content of which can be obtained through the IA API, as well

[2]See also the `randomForest R` package, which is based on Breiman's (2001) original implementation

[3]`https://data.webarchive.org.uk/opendata/ukwa.ds.2/`

[4]`https://archive.org/`

as the archival timestamp. We utilise two subsets of these data: the so-called Geoindex and the Host Link Graph. The first includes all the archived .uk webpages the web text of which contains at least one string in the form of a UK postcode, e.g. "B1 1AA", and we use this information to geolocated these wepbages, and the websites these webpages are contained within. The Geoindex dataset contains almost 700 million URLs which point to such webpages, the postcodes and the archival timestamp (Jackson 2017a). It should be highlighted that such a geolocation procedure using the HTML text and references to postcodes does not entail the same limitations with IP geolocation attempts (Zook 2000) and the 'here and now' issues linked to user generated data from social media (Crampton et al. 2013). These data have been employed before in answering social science research questions. Musso and Merletti (2016) reconstructed the UK business web ecosystem during the 1996-2001 period and Tranos, Kitsos, and Ortega-Argilés (2021) illustrated the long term regional productivity effects of the early adoption of web technologies. Tranos and Stich (2020) explored the role of online content of local interest in attracting individuals online and Hale et al. (2014b) mapped the web presence of the UK universities. However, to our knowledge this is the first time that such extended, but also granular in terms of space and time, archived web data has been utilised to model interregional flows and, more specifically, trade.

The Host Link Graph dataset was constructed by scanning the overall dataset for hyperlinks between websites (Jackson 2017b). In essence, this is a long edge list and each observation contains the website where the hyperlinks originate from, the website the hyperlinks point to, the number of hyperlinks between this origin-destination pair of websites and the archival timestamp. Between 2000 and 2010, 1.6 billions hyperlinks were found, with the majority being links between different webpages within the same website.

To create interregional flows of hyperlinks, we then combine the above datasets using the following workflow. Firstly, we aggregate the Geoindex data from webpages to websites by grouping together all archived webpages which are contained under the same website[5]. Because our aim is to predict interregional trade we only include in our analysis commercial websites by filtering out websites which are not part of the .co.uk second level domain (SLD) (Thelwall 2000). UK based companies are free to adopt a generic TLD such as .com and such websites are not included in our data. Nevertheless, we do not expect any substantial bias because of such omissions given the popularity of the .uk TLD. For instance, UK customers have a strong preference towards .uk websites for purchasing products and services (Hope 2017). Moreover, .co.uk is by far the most popular SLD under the .uk ccTLD (Tranos, Kitsos, and Ortega-Argilés 2021). Regarding the geolocation of such commercial websites, given that their mission is to support businesses (Blazquez and Domenech 2018), we expect that the self-reported physical addresses in the form of postcodes refer to trading instead of registration address. After all, "the firm must include on its website all the information it wants its real and potential clients to know, presenting it in the most adequate manner" (Hernández, Jiménez, and Martín 2009, p. 364).

During this aggregation process we keep track of how many unique postcodes are included in every website per year. Table 1 presents the frequencies of websites based on counts of unique postcodes per website for 2000. We create two subsets to test our models. We first train and test our models on data including only websites (and their hyperlinks), which contain only one unique postcode. Such websites may represent a

[5]For example the following webpages `http://www.examplewebsite.co.uk/webpage1` and `http://www.examplewebsite.co.uk/webpage2` are part of the `http://www.examplewebsite.co.uk/` website.

small company with a single trading location and, therefore, the website geolocation procedure may suffer from less noise. In 2000 72% of all archived websites included only one unique postcode. As a robustness check we then replicate the analysis for an extended subset of websites, which include up to 10 unique postcodes. We geolocate these websites by equally attaching them to multiple locations. This extended sample of websites includes 94% of all the archived websites in 2000.

Table 1. Second-level domain (SLD) names frequencies, 2000.

	level	freq	perc	cumfreq	cumperc
1	(0,1]	41,596	0.72	41,596	0.72
2	(1,2]	6,451	0.11	48,047	0.83
3	(2,10]	6,163	0.11	54,210	0.94
4	(10,100]	2,975	0.05	57,185	0.99
5	(100,1000]	646	0.01	57,831	1.00
6	(1000,10000]	62	0.001	57,893	1.00
7	(10000,100000]	4	0.0001	57,897	1

We then merge the geolocated website dataset with the hyperlinks edge list in order to create yearly edge lists with geolocated origin and destination websites. We clean these data by removing any website self-links and any possible duplicates. The final step is to aggregate these yearly postcode level edge lists into NUTS2 regional edge lists to match the interregional trade flow data.

Regarding trade data, we obtain the flows of imports and exports between the UK NUTS2 regions from the EUREGIO database (Thissen et al. 2018), 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. 2015) 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, Los et al. 2020).

Regional trade in the EUREGIO database is taken from the PBL Netherlands Environmental Assessment Agency egional trade data for the year 2000 as a prior to the estimations for the whole series 2000-2010 (Thissen, Diodato, and Van Oort 2013b, and Thissen, Diodato, and Van Oort (2013a)). This dataset was constructed by merging data from several sources: national accounts of the selected countries; international trade data on goods from (Feenstra et al. 2005) 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. 2018; Ivanova, Kancs, and Thissen 2019).

This database has been used recently in studies estimating the impacts of different economic shocks. Los et al. (2017), 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 sce-

nario. Similarly, Chen et al. (2018) used these data to estimate the regional exposure to Brexit for the whole European Union. Kitsos, Carrascal-Incera, and Ortega-Argilés (2019) employed these data to examine the role of local industrial embeddedness on economic resilience for the UK regions. Wilting et al. (2020), among others, estimated the subnational greenhouse gas and land-based biodiversity footprints in the EU regions using this database.

The descriptive statistics of the data we use to train and test our models, including the other three variables we employ – distance between the centroids of NUTS2 regions in the UK, employment and population density for NUTS2 regions – are reported in Table 2. These control variables (employment and population density) are included following gravity-type functions that are common when estimating economic flows between two geographical points (trade, migration, commuting, etc.). Such models normally use attraction factors (as the masses of the two regions) such as gross income in the region/country (Anderson and Van Wincoop (2003); Riddington, Gibson, and Anderson (2006)), or employment as a proxy when Gross Value Added (GVA) or GDP data is not available (Kimura and Lee (2006)). Also, employment by sector is often used in the estimation process of regionalisation of Input-Output models by the means of Location Quotients (Flegg and Webber (2000)). We choose employment because it is also available at a more disaggregated geographical level. Population density controls for the agglomeration of the regions complementing the employment variable in determining the economic size of the bodies (Greene (2013)). Distance works as a resistance effect. In addition, given the aim of the paper to predict interregional trade flows, we opted towards parsimony and, therefore, we tried to minimise the number of predictors.

Furthermore, Figure 1 plots the interregional flows for both hyperlinks and trade for the first and the last year of our study period. In 2000, hyperlinks were primarily concentrated in the South East of England. However, by 2010, we begin to see a similar pattern to the trade links, with the major flows still coming from the same regions, but with a wider coverage. We can observe an increase in the flow intensity between 2000 and 2010, but this is higher for the hyperlinks. In total, we have data for 1369 pairs between 37 NUTS2 regions for 11 years ($2000 - 2010$).

Table 2. Descriptive statistics

Statistic	N	Min	Mean	St. Dev.	Max
Hyperlinks	15,059	0	640.0	7,166.7	534,958
Distance (km)	15,059	0	265,735.1	164,759.7	779,523
Employment (000s of employees)	15,059	188.0	755.2	374.3	2,224.5
Population density (Hab. per km^2)	15,059	10.7	467.6	590.1	3,008.6

5. Results

The first step of the analysis involves training our RF models using data from years t and $t + 1$. Figure 2 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

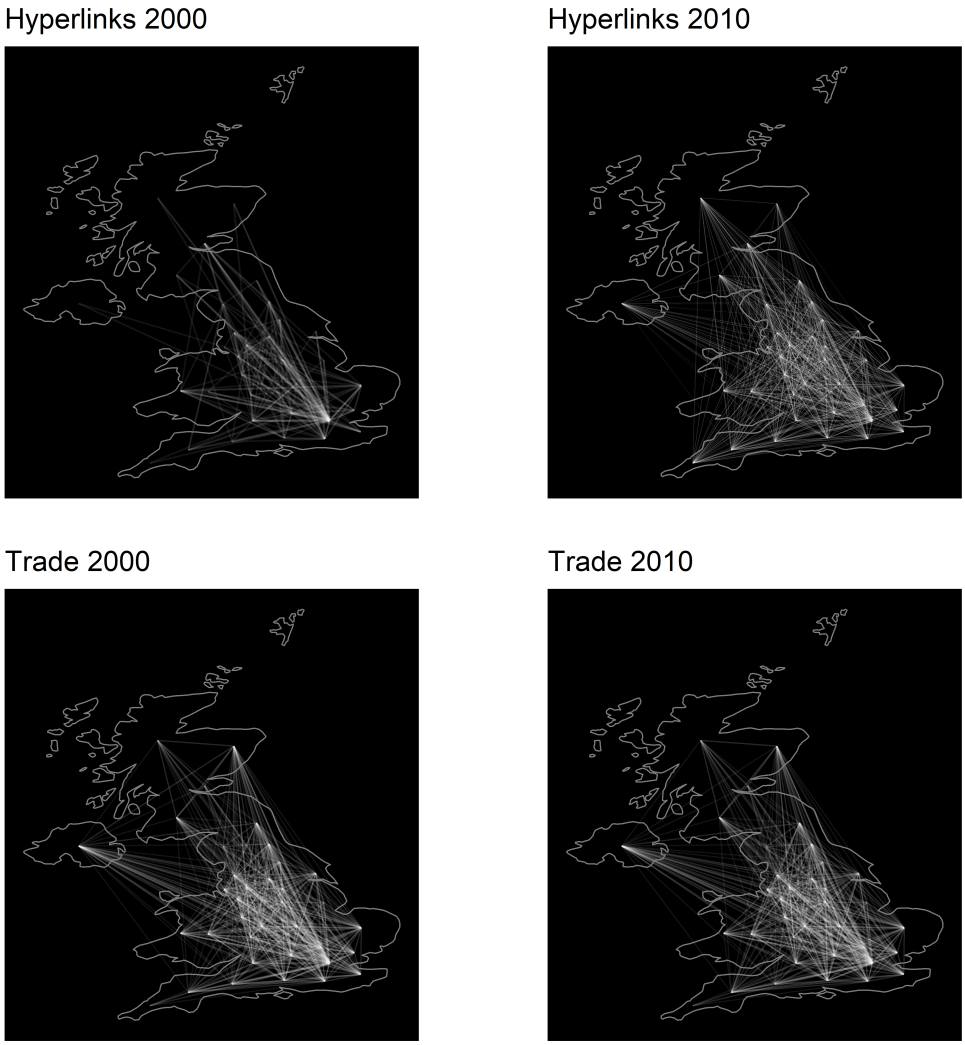

Figure 1. Distribution of Trade and Hyperlink flow data

the models estimated using data from years t and $t+1$ are applied on years $t+2$. In other words, we used the models trained with data from years t and $t+1$ and the explanatory variables for year $t+2$ to forecast interregional trade for year $t+2$. The yearly accuracy metrics are presented in Table 3 and the predicted versus the observed flows of interregional trade are plotted in Figure 3. Indeed, our models are able to make highly accurate temporal 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[6]. The drop of the predictive capacity of our model for 2010 can be attributed to the aftermath of the financial crisis and the use of data reflecting different business cycles – before and after the crisis. As Figure 3 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 temporal out of sample predictions.

Articles attempted to estimate subnational/national trade flows from a national/supranational Input-Output tables by means of Location Quotients techniques normally present larger error terms. For example, Pereira-López, Carrascal-Incera, and Fernández-Fernández (2020) using a novel method obtained errors for the domestic multipliers between 17.22 to 20.68, depending on the country studied. In a different study by Jiang, Dietzenbacher, and Los (2012), when estimating the Input-Output tables of the Chinese regions they obtain errors from 32.9 to 38.5 using traditional regionalisation methods. These errors are measured as Weighted Absolute Percentage Errors (WAPEs) (Sawyer and Miller (1983)). WAPEs express the absolute deviation in relation to the true value of each Input-Output coefficient. In other words, they report average error in percentage terms (Lamonica and Chelli (2018); Pereira-López, Carrascal-Incera, and Fernández-Fernández (2020)). In both cases these estimations were done for separate economies (single-table), which means that they were estimating intraregional (or intracountry) sectoral trade flows (but bilateral sector flows in any case).

Table 3. Accuracy metrics for predicting 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 4 reports the R-squared values for the predictions of the unseen $t+2$ interregional trade flows for each sector. In general, our models achieve higher accuracy in trade of goods (s1-s8) than services (s10-s15), which are conventionally considered as non-tradable (Jensen et al. 2005). We can also observe a drop in accuracy for service

[6]As a comparison, the Appendix provides the same results based on a LASSO estimator, which are inferior to the ones acquired through RF.

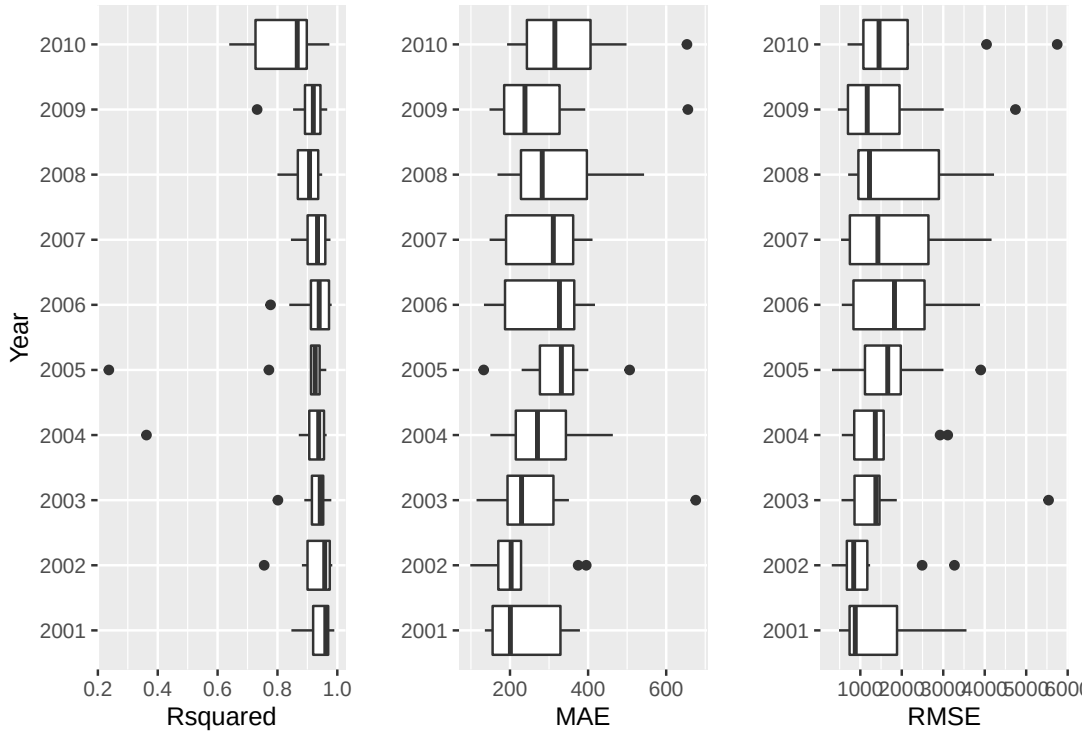

Figure 2. Accuracy metrics

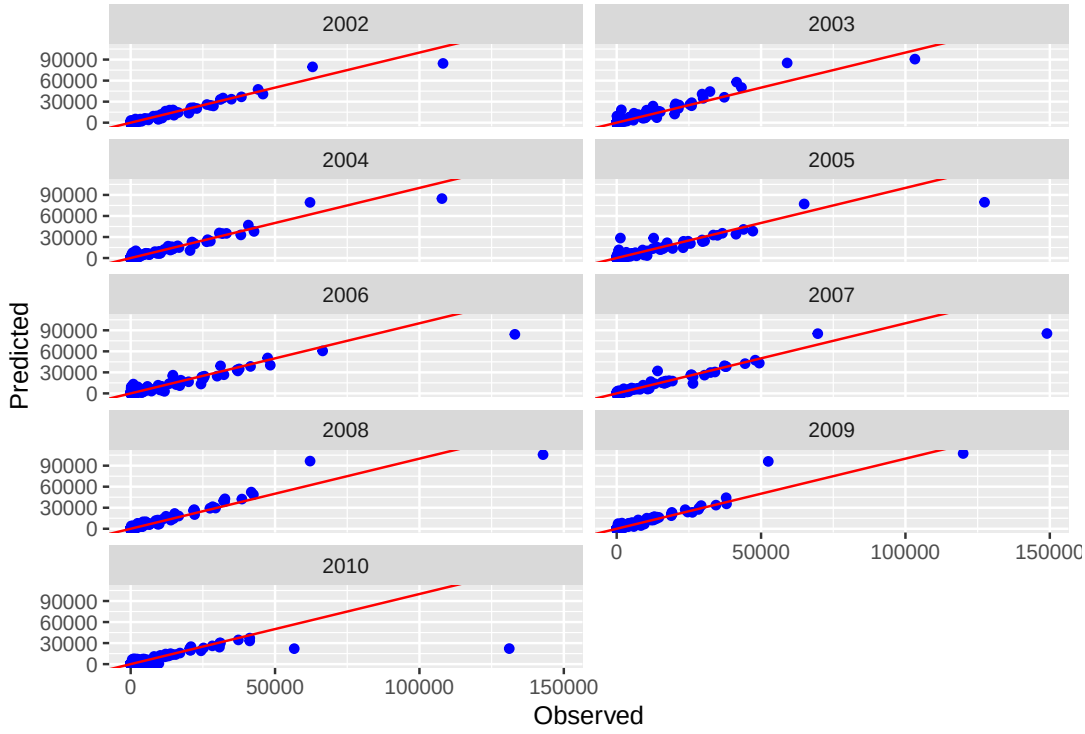

Figure 3. Predicted vs. observed interregional trade by year

sectors in 2010 due to the financial crisis and the related knock on effects. The decrease of interregional trade volume makes it more difficult to predict. As expected, our model performs worse in some specific sectors. The most obvious example is hospitality (s11), for which the R-squared between the predicted and observed values for 2004 and 2010 drops below 0.25 and it does not exceed 0.75 for the whole study period. This can be attributed to the strong local and intraregional trade dependencies of this sector. 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 0.75. Our models perform exceptionally well in predicting unseen interregional trade flows regarding manufacturing and construction as well as equipment.

In summary, our results show a clear pattern of tradable sectors vs. non-tradable sectors. Even though references such as Gervais and Jensen (2019) and Jensen et al. (2005) challenge this conventional view of goods as tradable and services as non-tradable, in Gervais and Jensen (2019) for the US they find that Manufacturing products are 75 tradable and 25 non-tradable (S3 to S8 sectors in our study), Recreation and Food services (the most comparable one to our Hospitality sector) is 86 non-tradable, and Real Estate and Leasing is 79 non tradable, among other results they obtain.

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 by excluding (1) the distance and (2) 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 5, which also includes the metrics for the base models presented in Table 3 for direct comparison. The main message from Figure 5 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. Two are the key messages from Figure 5. Firstly, achieving R-squared values of up to 0.86 without using a physical distance feature, which has traditionally been the main explanatory variable of bilateral trade, is indicative of the predictive power of our hyperlinks approach. Secondly, the gap in terms of the prediction accuracy between the models with and without distance decreases over time. This illustrates that over time, as the adoption rate of web technologies increased, interregional trade flows left more 'digital breadcrumbs' behind and, therefore, are better reflected in the volumes of interregional hyperlinks (Rabari and Storper 2014). Nevertheless, the predictive capacity of distance remains unchallenged at large as the green lines in Figure 5 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 10 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 4 and the predicted versus the observed interregional trade flows, which are plotted in Figure 6. Indeed, R-squared drops below 0.90 for only two years (2009 and 2010). Again, the largest prediction errors are linked to the regional pairs with the highest volume of regional trade – that is London's

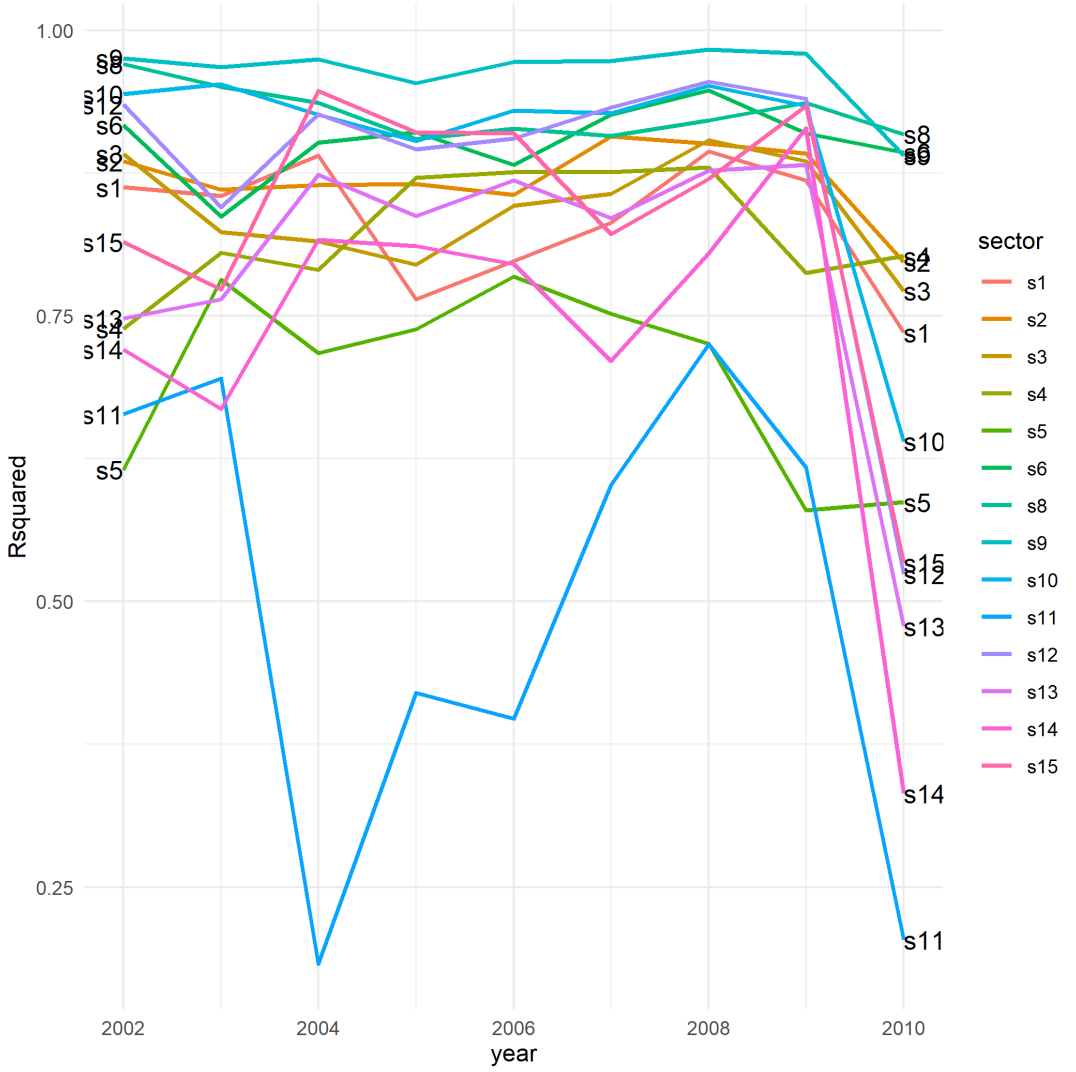

Figure 4. R-squared for t + 2 out of sample predictions per sector.
Notes: s1: Agriculture, s2: Mining, s3: Food, s4: Textiles, s5: Chemicals, s6: Equipment, s8: Manufacturing; s9: Construction, s10: Distribution, s11: Hospitality, s12: Transport, s13: Financial, s14: Real Estate, s15: Non-Market Services.

intraregional trade.

Table 4. 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. A local level application

The final step in our workflow is to utilise the NUTS2 models we trained and validated for a more spatially disaggregated application. This exercise illustrates the value of our research framework in mapping interregional trade flows at adissagregated spatial level, at which usually such data are not collected. Specifically, we take advantage of the point nature of our hyperlink data and we aggregate them at the Local Authority District (LAD) level. This is the main subnational administrative division in the UK. We then apply the NUTS2 model, which was trained on data from 2008 and 2009 and tested on 2010 trade data, on the LAD hyperlinks, employment, population density and distance data and we make predictions regarding LAD to LAD trade flows for 2010, the most recent data. Although we are not able to validate these predictions at the LAD level as there is no trade data available at this level, this exercise provides a unique opportunity to map trade flows at this level of spatial granularity.

We briefly discuss here the predictions for two LAD: Camden in central London and Birmingham , the second biggest city in the UK. We present the import and export trade flows for these LAD in Figure 7 [7]. While both of these examples illustrate the importance of distance in trade flows – light colour lines are concentrated near Camden and Birmingham – Camden appears to have more light colour links not only with adjacent LAD, but also with more distant ones both in terms of imports and exports. Not surprisingly, Camden's reach appears to be more extended than Birmingham's.

The above example illustrates the capacity of our research framework for spatially disaggregated analysis of trade flows. To our knowledge, observed trade data at this level is very hard to be found. Importantly though, the LAD level in the UK represents administrative units and, therefore, our framework can be utilised by local authorities to design relevant local policies. Without these data and a modelling framework at this geographical level, it is nearly impossible to do accurate ex ante evaluations of place-based policies. With these predictions, any LAD could calculate the net impact of a policy in their region and estimate the possible benefits for the residents.

[7]The online appendix provides the URL to interactive visualisations of the hypelink flows at the NUTS2 level and the trade flow predictions for LADs

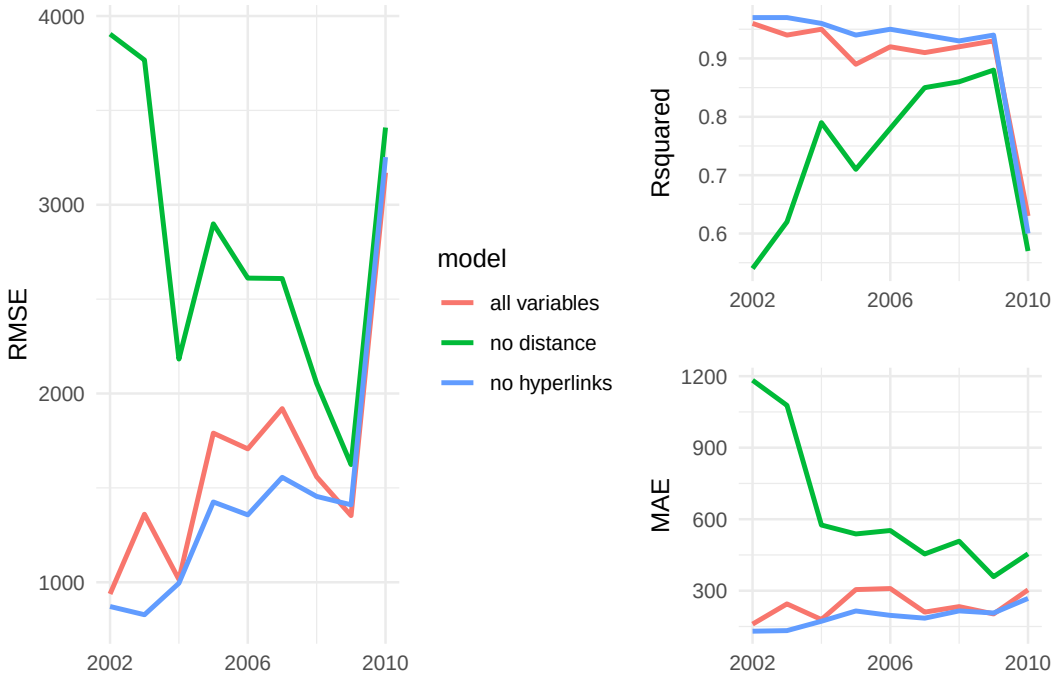

Figure 5. Accuracy metrics for alternative specifications

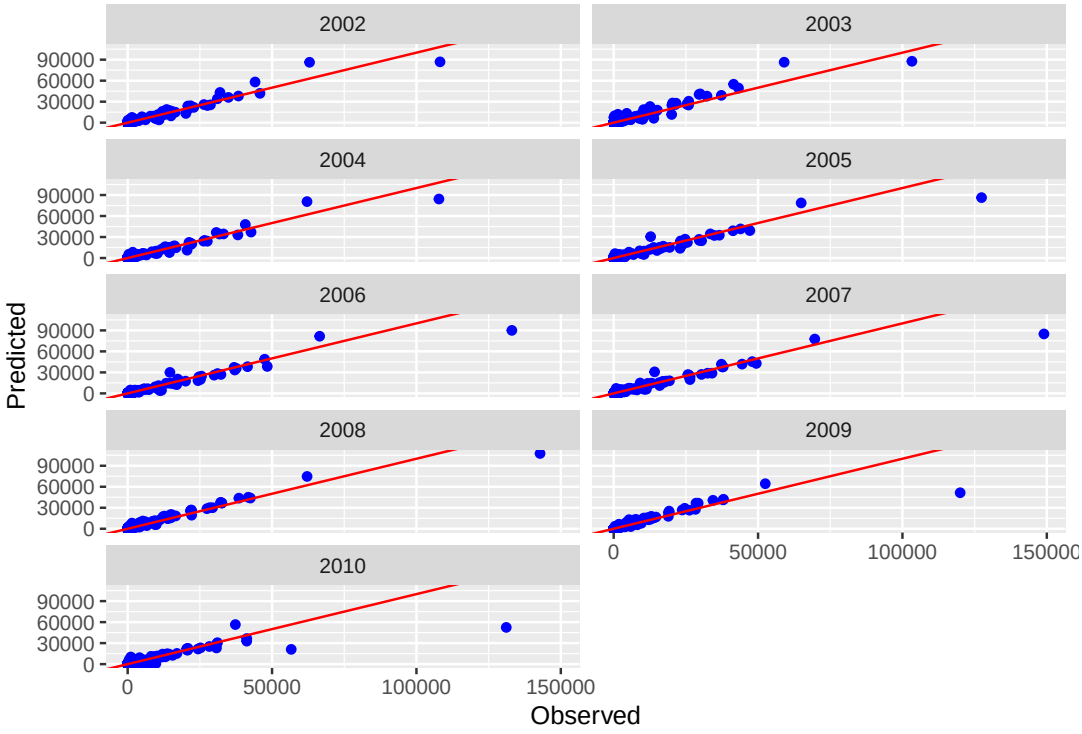

Figure 6. Predicted vs. observed interregional trade by year for multiple postcodes

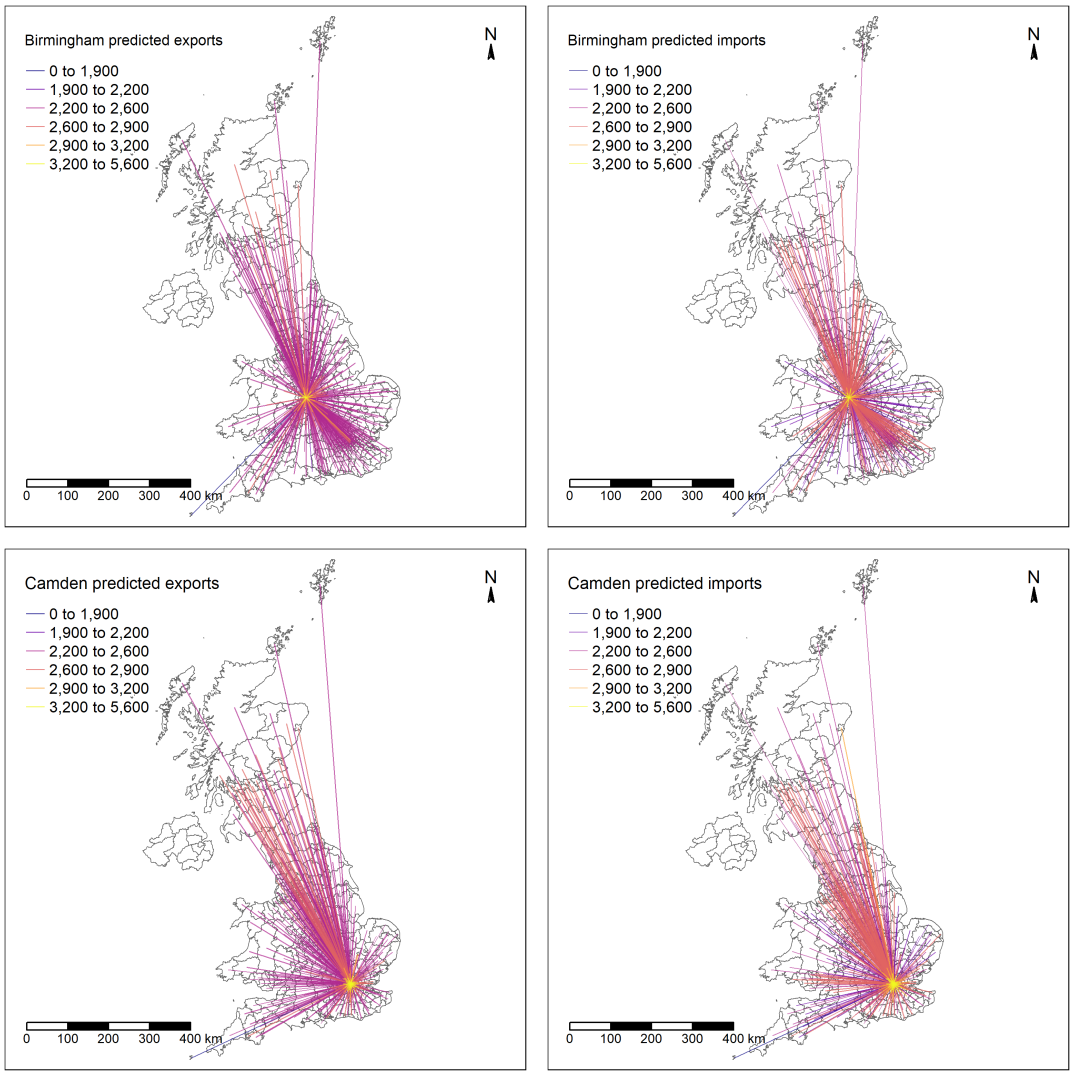

Figure 7. Local Authority Districts trade predicitions: top left- Birmingham Exports, top right - Birmingham imports, bottom left - Camden exports, bottom right - Camden imports

7. Conclusions

Despite the complexity of interregional trade and its importance in regional economic performance and, consequently, regional economic policies, it is well established in the literature that interregional trade is difficult to be observed. This is because such data are mostly available at a country level and hardly ever are monitored at a local or regional level. The current state of the art of empirical studies focusing on modelling international trade tend to be explanatory in their nature and are mostly based on pure distance decay measures.

Our paper aims to address this gap by proposing an innovative research framework, which is based on openly accessible web data and predictive models. On top of using variables such as distance, employment and population density, we employ web data regarding the number of hyperlinks between geolocated commercial websites aggregated at the regional level during the 2000-2010 period. These data, in essence, represent the digital breadcrumbs that trade leaves behind nowadays and we utilise them in order to predict interregional trade flows. By building a rolling forecasting workflow based on RF we are able achieve very accurate out of sample temporal predictions of interregioanal trade flows in the UK. We are able to also disaggregate our models at a sectoral level and illustrate the sectors for which our models perform better. This sectoral variation in our predictive capacity matches our expectations, which are routed in the relevant literature. We also perform different sensitivity tests, which further reinforce the value of our framework. Finally, we utilise our regional models to produced spatially disaggregated trade flows at a local level – the UK LAD. This is of particular importance as this is the main subnational administrative authority in the UK and such illustrations can be helpful to design local economic policies.

The current wide availability of archived data makes our framework easily applicable to different temporal and geographical contexts. Hence, using more recent archived web data, something which of course involves upfront investment regarding developing the necessary software infrastructure, will allow the nowcasting of interregional trade data for a variety of local scales to much administrative authorities responsible for designing local and regional economic policies. Identifying these types of external dependencies and vulnerabilities to supply chain disruptions is essential for the regional wellbeing, as the recent Covid-19 crisis has shown us. This can help us to anticipate local exposures and knock-on effects to shocks and to elaborate mitigating policies almost in real time.

References

Ainsworth, Scott G, Ahmed Alsum, Hany SalahEldeen, Michele C Weigle, and Michael L Nelson. 2011. "How much of the web is archived?" In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, 133–136. ACM.

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

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

Arto, Iñaki, 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): 327–353.

Athey, Susan, and Guido W Imbens. 2019. "Machine learning methods that economists should know about." *Annual Review of Economics* 11: 685–725.

Barca, Fabrizio. 2009. "An Agenda for a Reformed Cohesion Policy-Independent Report."

European Commission, Brussels .

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

Blazquez, Desamparados, and Josep Domenech. 2018. "Big Data sources and methods for social and economic analyses." *Technological Forecasting and Social Change* 130: 99–113.

Boero, Riccardo, 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): 225–238.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45 (1): 5–32.

Caruana, Rich, 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*, ICML '08, New York, NY, USA, 96–103. Association for Computing Machinery. https://doi.org/10.1145/1390156.1390169.

Chen, Wen, 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): 25–54.

Chun, Yongwan, 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): 583–591.

Chung, Joo. 2011. "The geography of global internet hyperlink networks and cultural content analysis." PhD diss., Dissertation, University at Buffalo.

Crampton, Jeremy W., Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook. 2013. "Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb." *Cartography and Geographic Information Science* 40 (2): 130–139. https://doi.org/10.1080/15230406.2013.777137.

Credit, Kevin. 2021. "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* .

David, H, David Dorn, and Gordon H Hanson. 2013. "The China syndrome: Local labor market effects of import competition in the United States." *American Economic Review* 103 (6): 2121–68.

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

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

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

Dietzenbacher, Erik, 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): 71–98.

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

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

Fingleton, Bernard, Harry Garretsen, and Ron Martin. 2012. "Recessionary shocks and regional employment: evidence on the resilience of UK regions." *Journal of regional science* 52 (1): 109–133.

Flegg, Anthony T, and CD Webber. 2000. "Regional size, regional specialization and the FLQ formula." *Regional Studies* 34 (6): 563–569.

Gervais, Antoine, and J Bradford Jensen. 2019. "The tradability of services: Geographic concentration and trade costs." *Journal of International Economics* 118: 331–350.

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

Greene, William. 2013. "Export potential for US advanced technology goods to India using a gravity model approach." *US International Trade Commission, Working Paper* (2013-03B): 1–43.

Guan, Dabo, Daoping Wang, Stephane Hallegatte, Steven J Davis, Jingwen Huo, Shuping Li, Yangchun Bai, et al. 2020. "Global supply-chain effects of COVID-19 control measures." *Nature human behaviour* 4 (6): 577–587.

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

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

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

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

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

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

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

Hernández, Blanca, Julio Jiménez, and M José Martín. 2009. "Improved estimation of regional input–output tables using cross-regional methods." *International Journal of information management* 29 (5): 362–371.

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

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

Holzmann, Helge, Wolfgang Nejdl, and Avishek Anand. 2016. "The Dawn of today's popular domains: A study of the archived German Web over 18 years." In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference*, 73–82. IEEE.

Hope, Oli. 2017. "The changing face of the online world." https://www.nominet.uk/changing-face-online-world/. Accessed: 2021-03-05.

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

Isard, Walter. 1951. "Interregional and regional input-output analysis: a model of a space-economy." *The review of Economics and Statistics* 33 (4): 318–328.

Isard, Walter. 1956. "Location and space-economy." .

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

Jackson, Andrew N. 2017a. "JISC UK Web Domain Dataset (1996-2010) Geoindex." .

Jackson, Andrew N. 2017b. "JISC UK Web Domain Dataset (1996-2010) Host Link Graph." .

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

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

Jensen, J. Bradford, Lori G. Kletzer, Jared Bernstein, and Robert C. Feenstra. 2005. "Tradable Services: Understanding the Scope and Impact of Services Offshoring [with Comments and Discussion]." *Brookings Trade Forum* 75–133. `http://www.jstor.org/stable/25058763`.

Jiang, Xuemei, Erik Dietzenbacher, and Bart Los. 2012. "Improved estimation of regional

input–output tables using cross-regional methods." *Regional Studies* 46 (5): 621–637.
JISC, and the Internet Archive. 2013. "JISC UK Web Domain Dataset (1996-2013)." *The British Library* .
Jones, Brant W, 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): 99–111.
Keßler, Carsten. 2017. "Extracting central places from the link structure in Wikipedia." *Transactions in GIS* 21 (3): 488–502.
Kimura, Fukunari, and Hyun-Hoon Lee. 2006. "The gravity equation in international trade in services." *Review of world economics* 142 (1): 92–121.
Kitsos, Anastasios, André Carrascal-Incera, and Raquel Ortega-Argilés. 2019. "The role of embeddedness on regional economic resilience: Evidence from the UK." *Sustainability* 11 (14): 3800.
Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction policy problems." *American Economic Review* 105 (5): 491–95.
Krüger, Miriam, 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).
Kuhn, Max, et al. 2008. "Building predictive models in R using the caret package." *Journal of statistical software* 28 (5): 1–26.
Lamonica, Giuseppe Ricciardo, and Francesco Maria Chelli. 2018. "The performance of non-survey techniques for constructing sub-territorial input-output tables." *Papers in Regional Science* 97 (4): 1169–1202.
Last, Mark, Oded Maimon, and Einat Minkov. 2002. "Improving stability of decision trees." *International Journal of Pattern Recognition and Artificial Intelligence* 16 (02): 145–159.
Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, et al. 2009. "Social science. Computational social science." *Science (New York, NY)* 323 (5915): 721–723.
Leamer, E., and R. Stern. 1971. "Quantitative International Economics." *Journal of International Economics* 1: 359–361.
Liaw, Andy, Matthew Wiener, et al. 2002. "Classification and regression by randomForest." *R news* 2 (3): 18–22.
Lima, Marcio Salles Melo, and Dursun Delen. 2020. "Predicting and explaining corruption across countries: A machine learning approach." *Government Information Quarterly* 37 (1): 101407.
Lin, Jia, Alexander Halavais, and Bin Zhang. 2007. "The blog network in America: blogs as indicators of relationships among US cities." *Connections* 27 (2): 15–23.
Linnemann, Hans. 1966. *An econometric study of international trade flows*. North-Holland Pub. Co.
Los, Bart, Philip McCann, John Springford, and Mark Thissen. 2017. "The mismatch between local voting and the local economic consequences of Brexit." *Regional Studies* 51 (5): 786–799.
Los, Bart, 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): 66–92.
Los, Bart, Marcel P Timmer, and Gaaitzen J de Vries. 2016. "Tracing value-added and double counting in gross exports: comment." *American Economic Review* 106 (7): 1958–66.
Matter, Regions. 2009. "Economic Recovery." *Innovation and Sustainable Growth.-Paris: OECD* .
McCann, Philip, and Raquel Ortega-Argilés. 2015. "Smart specialization, regional growth and applications to European Union cohesion policy." *Regional studies* 49 (8): 1291–1302.
Meijers, Evert, and Antoine Peris. 2019. "Using toponym co-occurrences to measure relationships between places: review, application and evaluation." *International Journal of Urban Sciences* 23 (2): 246–268.

Mikkonen, Kauko, and Martti Luoma. 1999. "The parameters of the gravity model are changing–how and why?" *Journal of Transport Geography* 7 (4): 277–283.

Miller, Ronald E, and Peter D Blair. 2009. *Input-output analysis: foundations and extensions*. Cambridge university press.

Mozolin, Mikhail, J-C Thill, and E Lynn Usery. 2000. "Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation." *Transportation Research Part B: Methodological* 34 (1): 53–73.

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

Musso, Marta, and Francesco Merletti. 2016. "This is the future: A reconstruction of the UK business web space (1996–2001)." *New media & society* 18 (7): 1120–1142.

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

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

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

Oshan, Taylor M. 2020a. "Potential and pitfalls of big transport data for spatial interaction models of urban mobility." *The Professional Geographer* 72 (4): 468–480.

Oshan, Taylor M. 2020b. "The spatial structure debate in spatial interaction modeling: 50 years on." *Progress in Human Geography* 0309132520968134.

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

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

Pereira-López, Xesús, André Carrascal-Incera, and Melchor Fernández-Fernández. 2020. "A bidimensional reformulation of location quotients for generating input–output tables." *Spatial Economic Analysis* 15 (4): 476–493.

Pontius, Robert Gilmore, 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): 111–142.

Pourebrahim, Nastaran, Selima Sultana, Amirreza Niakanlahiji, and Jean-Claude Thill. 2019. "Trip distribution modeling with Twitter data." *Computers, Environment and Urban Systems* 77: 101354.

Rabari, Chirag, 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): 27–42.

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

Riddington, Geoff, Hervey Gibson, and John Anderson. 2006. "Comparison of gravity model, survey and location quotient-based local area tables and multipliers." *Regional Studies* 40 (9): 1069–1081.

Salvini, Marco M, 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): 228–248.

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

Sawyer, Charles H, and Ronald E Miller. 1983. "Experiments in Regionalization of a National Input—Output Table." *Environment and Planning A* 15 (11): 1501–1520.

Serrano, Ma Ángeles, and Marián Boguñá. 2003. "Topology of the world trade web." *Phys. Rev. E* 68: 015101. https://link.aps.org/doi/10.1103/PhysRevE.68.015101.

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

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

Sinha, Parmanand, 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: 132–145.

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

Thelwall, Mike. 2000. "Who is using the .co.uk domain? Professional and media adoption of the web." *International Journal of Information Management* 20 (6): 441–453. https://www.sciencedirect.com/science/article/pii/S0268401200000384.

Thelwall, Mike. 2002a. "Evidence for the existence of geographic trends in university web site interlinking." *Journal of Documentation* .

Thelwall, Mike. 2002b. "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): 483–491.

Thelwall, Mike, and Liwen Vaughan. 2004. "A fair history of the Web? Examining country balance in the Internet Archive." *Library & information science research* 26 (2): 162–176.

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

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

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

Thissen, Mark, 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): 159–180.

Thissen, Mark, 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." .

Timmer, Marcel P, 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): 575–605.

Tinbergen, Jan. 1962. "Shaping the World Economy The Twentieth Century Fund." *New York* 330.

Többen, Johannes, and Tobias Heinrich Kronenberg. 2015. "Construction of multi-regional input–output tables using the CHARM method." *Economic systems research* 27 (4): 487–507.

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

Tranos, Emmanouil, and Chrisotph Stich. 2020. "Individual internet usage and the availability of online content of local interest: A multilevel approach." *Computers, Environment and Urban Systems* 79: 101371.

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

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

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

Wang, Lingjing, 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).

Wilting, Harry C, 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* .

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

Zook, Matthew A. 2000. "The web of production: the economic geography of commercial Internet content production in the United States." *Environment and planning A* 32 (3): 411–426.