Editor comments

I have now received all reviews for your paper. Based on the comments, I ask that you undertake Moderate Revisions and resubmit the revised manuscript (via Manuscript Central) for further consideration. Comments from reviewers are pasted to the end of this letter. I would like to echo the suggestions made by Reviewer #1, #2, and #4 that in the revision, please clarify and sharpen the key contribution of the work as random forest is already a 'bread and butter' method (as put by reviewer #4).

Thank you for offering the opportunity to revise and resubmit our paper. In this revised version of the manuscript, we have amended the analysis and text to address the very useful referee comments. We made sure to clarify and sharpen the key contribution of the paper. We do indeed use an estimator which is 'bread and butter' for data science, but we believe that our research framework is novel as we utilise openly available web data to make *out of sample predictions* instead of estimating a traditional, explanatory, parametric model. Importantly, these predictions are useful for both for researchers and policy makers as there are very limited data regarding regional trade flows and hardly any for more granular aerial units. The current version of the paper also offers an online appendix which visualise these disaggregated predictions.

Below we provide detailed answers to the specific comments.

Reviewer: 1

This manuscript reports a work of predicting interregional trade by training random forests (RF) with particular features aggregated web data, that is, interregional network of hyperlinks between geolocated commercial webpages (with ".com.uk" domain and geolocated by UK postcode stated in webpages). The time range of the webpage records for this study was 2000-2010, which were split as biyearly data for training RF and the third year for prediction. RF with such features derived from web data permits a finer scale prediction of not only all companies but also individual industrial sectors, compared with the conventional used survey data having much coarser spatial and temporal resolution. Sensitivity analysis on several factors were also conducted and discussed to show the good performance of RF. The workflow is reasonable and reported detailedly in a replicable way.

We want to sincerely thank the referee for their report. Their points have helped us to increase the value of the paper. Thank you also for highlighting the qualities of our workflow.

The main idea in this study is to use web data to fix the lack of regional trade data (especially with fine scale, UK NUTS2 regions). I'm not expertise of economic geography. My general comment is below. New data (for application domain)? Authors did a good job on data clean and aggregation.

Thank you for recognising our data cleaning and aggregation processes.

However, from scientific perspective, aggregation of webpages for this domain is comparatively new but not quite innovative (as review in Page 4 of this manuscript).

Yes, we are not claiming that we are the first ones utilising web data to answer social science research question. As you noted we reviewed the literature which utilised similar data. But we still think that our approach is novel as never before archived web data have been employed in that context and extend. Hence, we added the following sentence in Section 4, p. 9:

> 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.

New method? RF is not new and has been widely used in geography (including socio-economic domain).

Point taken. What we wanted to emphasise was our aim to make out of sample predictions instead of estimating a traditional, explanatory, parametric model. Indeed, RF are well established estimators. We copy here the amended text from the Introduction:

> Our paper contributes to this line of inquiry by utilising novel web data and machine learning algorithms to make out-of-sample predictions for the UK NUTS2 regions during the period 2000-2010.
>
> 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.

and the Conclusions:

> Our paper aims to address this gap by proposing an innovative research framework, which is based on openly accessible web data and predictive models.

New finding? When dataset used in this study was 2000-2010, the drop in 2010 due to the financial crisis for almost all sectors (not only the service sector mentioned in the manuscript) as shown in Figure 4 cannot be further discussed in details on the performance of the proposed method, such as the model performance consistency before and after the crisis. It's a pity. Such a current limit also makes the end statement of this manuscript (potentials related to "recent COVID-19 crisis", and "help us to anticipate local exposures and knock-on effects to shocks") to be much less supported. Note that this topic may also have been explored under the COVID-19 pandemic event. If authors may further add discussion on this point, it should be considered to relate the discussion with that.

We agree that the period of the database limits the conclusions about the method during crises. At the end, we are training the model with past data, but the statement about the usefulness of the results obtained is made regarding the utility of predicting interregional trade flows and the granularity of these predictions. In other words, the proposed research framework allows to obtain predictions in order to track the propagation processes linked with positive or negative economic shocks. Intra and interregional trade data are very scarce (even scarcer or nonexistent at the local authority/municipal level) and having a research framework that provides some insights is better than trying to develop a simulation model to understand and estimate the regional economic exposure with just national survey data obtained every 5 years. Economic exposure analyses to negative shocks (such as the Covid-19 or the financial crisis) need to be done through simulations using data just before the arrival of a crisis, which is not the same as predicting the trade flows in a crisis year. The Covid-19 reference in the conclusions is not related to the potential capabilities of the method proposed, but to put into consideration the necessity to highlight the importance of "identifying external dependencies and vulnerabilities to supply chain disruptions".

Other specific comments are below:

Section "4 Data": Please make it clear if a web page in Geoindex includes a UK postcode string not only in this webpage (might be other unrelated information) but also in particular position for stating the location of this webpage host. Otherwise, it may likely be a misinformation for geolocating the company of the webpage.

If we understand the above comment correctly, the reviewer is asking to elaborate on the geolocation process. The raw data (Geoindex) include the unique postcodes mentioned in each archived webpage. Our aggregation process allows us to assess this information not at the webpage level, but instead at the website level and, therefore, we use these terms carefully in the paper. If it was the former, the reviewer would have been right to question what the inclusion of a postcode actually means in a single webpage. However, because of the aggregation process we observe websites with a range of postcodes from one unique postcode to thousands as per Table 1. Our expectation is that the postcodes included in websites with only one unique postcode represent the trading addresses of these commercial activities. We believe that this is a fair expectation as we know from the literature that commercial websites perform specific missions. See for instance the below amended text in p. 9:

> 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 et al., 2009: 364).

The following text is also included in Section 4, which further explains this process:

> we use this information to geolocated these wepbages and the websites these webpages are contained within.
>
> Firstly, we aggregate the Geoindex data from webpages to websites by grouping together all archived webpages which are contained under the same website[5].

[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.

We firstly analyse websites with one unique postcode included in all their archived webpages. As per Blazquez and Domenech (2018) we expect these websites to represent economic activities trading in the unique location included in the archived webpages. Then, as a robustness check we repeat our analysis for websites with up to 10 unique postcodes and the results remain at large the same. Considering all the above, we are confident that our geolocation process is meaningful and informative.

Besides those from web data, other input features for RF (i.e., employment, and population density) used in this study have not been justified on the selection. Why these two? And why not to include other potential features?

We have now added the following paragraph in the Data section in p. 11:

> 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 et al., 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 regionalization 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.

Section 6. Why this part used 2010 data for test, when the former tests showed clearly that the prediction for 2010 was the worst distinctly.

As are aiming to illustrate the value of our research framework, we opted towards using the latest available data. Of course, there is a trade off as this time period is affected by the financial crisis. We do believe though that Section 6 still illustrates the applicability of our framework to more granular scales, for which there is no other available data. If the reviewer thinks that we need to use earlier data for this illustration, we are happy to do so.

Page 9: "78% of all archived websites included only one unique postcode" - ? 72% according to Table 1.

The typo has now been corrected.

Table 2: unit of row? Max of "Employment" is 2224.5 - is it reasonable for ".5" (for counting part-time employments)?

Regarding employment, it is expressed in thousands, which is why we have some ".5" values. The units have now been added in Table 2.

Page 12: “obtain errors from 32.9 to 38.5 using traditional regionalisation method” - unit for the error? Some other places with the similar question.

We have now added to following text in the results section (p. 13) to clarify this point:

> These errors are measured as Weighted Absolute Percentage Errors (WAPEs). 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 et al., 2020). In that sense, the R-squared is a comparable measure.

Figure 5: the x-axis is too crowded.

Figure 5 has now been amended.

Some minor spelling errors: “ccLTD” in Page 9, “interregional trafde” in Page 17.

The typos have now been corrected.

Reviewer: 2

The paper proposes a novel methodological framework based on Machine Learning (ML) and in particular the Random Forest (RF) technique in order to predict interregional trade among UK regions. The results focus on total trade flows and by sector and are validated with a series of tests and using the EUREGIO regional Input-Output database and show a high degree of predictive capacity for the technique. The author(s) provide an illustrative example of how the methodology can provide estimates of inter-local trade flows for areas for which there is little to none such information. The paper is novel, interesting and topical in the sense that it contributes to the growing need for data at a variety of spatial scales using data science techniques. I have two main comments below, followed from some more detailed points and suggestions.

We want to thank the referee for their valuable comments and also for recognising the novelty and the contribution of our paper.

\1) Contribution of the paper. The author(s) do a good job in highlighting the methodological contribution of the paper, but I believe they undersell the paper in terms of its contribution to policy. I suggest that changes throughout the document should highlight the value of the method for regional research and policy which at the moment is saved for the last part of the paper.

Following the suggestion of the referees, we have rewritten parts of the introduction to highlight its contribution also in terms of policy. Specifically, we have added the following text in the Introduction section:

> 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.
>
> 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.

\2) Structure and detail. The contribution and intuition of the paper is often obscured by the technical discussion that is dominant at the moment. Below are some more detailed comments and suggestions in making more prominent what the problem is, what the proposed solution is and how it contributes. Phrasing these in non-technical terms, I believe will expand the target audience of the paper.

- Abstract

A lot of the abstract is spent on the JISC UK dataset and the construction of the extracted data instead of the importance of the framework.

There is also little mention as to how the predictive capability is tested.

A revision of the phrasing with more direct language would improve the sharpness of the abstract. i.e. there is a mixture of third and first person and the last sentence is hard to understand.

The abstract has now been re-written:

> 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.

- Introduction

In keeping with the abstract, I find the introduction somewhat underdeveloped in terms of highlighting the importance of what is proposed. As a reader, I would be interested to see more of what are we missing from not knowing the knowledge that this paper contributes. The first two-three paragraphs could do more to generate the intuition and need for the paper.

As noted above, we have restructured parts of the introduction. We also added the following text at the end of paragraph 2:

> 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.

It would be helpful to add somewhere in the 4th or 5th paragraph, a sentence or two on what the RF model does and what it allows you to do and why it may be superior.

This is indeed very important, but we think we explicitly state it in paragraph 6 of the introduction:

> 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.

To make space for the above changes, I believe the penultimate paragraph that discusses how extensive the use of RF is, to be moved to the literature review or methodological framework section and be replaced in the introduction with a couple of reference and a sentence. As long as there is discussion on how the method allows you to tackle specific issues, there is little need in discussing how extensively it has been used in the introduction.

We have now move this paragraph to Section 3 – second paragraph of the section.

- Literature overview

The overview is mature and well structured with reference to the available (or not) interregional data and the use of webometrics. I would suggest it is missing an opening paragraph that highlights what we miss by not having insight into interregional trade flows.

Thank you for your nice comment. We have now added the following opening paragraph.

> 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.

- Methodological Framework

I believe there is need for more intuition on why the RF methodology is needed right at the beginning of the section. What is the main problem the paper addresses? What are the data attributes that the RF is best placed to utilise and provide the answers we need?

The opening paragraph of the third section detailed the attributes of RF, but indeed, it did not link them to our problem. We have added the following sentence at the end of the first paragraph of Section 3.

> All the above advocate towards utilising RF in this paper as we aim to do out-of-sample predictions of data, which are skewed and have outliers (see for instance Figure 3.

- Data

Table 2 provides descriptive statistics for distance but the variable has not been discussed so far?

Yes, this was an omission. We have now added the following in the last paragraph of Section 4:

> 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.

We have also added the following text regarding the other predictors included in our model (p. 11):

> 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 et al., 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 regionalization 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.

It would be useful to clarify how many regions are considered and what makes up the 15059 observations in Table 2

The below sentence has now been added at the end of Section 4:

> In total, we have data for 1369 pairs between 37 NUTS2 regions for 11 years (2000-2010).

- Results

There is little explanation of the reduction in Rsquared in Table 3 for year 2010. Could it be due to using data from a different stage in the business cycle? This is mentioned later on in passing on the results for services.

Yes, that is exactly the point. Using previous years' information for predicting 2010 has some drawbacks reflected in the lower Rsquared observed. Just with data on distance, employment, population density and hyperlinks of a different stage in the business cycle is challenging to accurately predict 2010 trade flows, as signalled by the referee. We opted towards a parsimonious model that makes the interpretation of the results straightforward.

We added the following text to provide further explanation when we report the main results in p. 13:

> 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.

Also it would be useful to understand better the comparability between Pereira-Lopez et al (2020) & Jiang et al (2012) to this study and its accuracy metrics to support the argument that similar articles using different methods have higher error terms.

Both in Jiang et al., (2012) and in Pereira-Lopez et al., (2020) they measure the accuracy of their methods using the Weighted Absolute Percentage Errors (WAPEs). WAPEs express the absolute deviation in relation to the value of the true value of each Input-Output coefficient. In other words, it tells us the average error in percentage terms (Lamonica and Chelli, 2018; Pereira et al., 2020). We have included this in the results section in p. 13, as suggested.

It would be useful to add some references in explaining where the accuracy metrics are lower such as the claim on tourism trade dependencies at the top of p. 14. Similarly for real estate.

Here the point is that sectors that don't have many commercial links with other sectors are going to be always more difficult to predict than those that are buying and selling their products to other sectors more often. The results show a clear pattern of tradable vs. non-tradable sectors. Overall, what can be seen in Figure 4 is that services that are conventionally labour-intense (s11, s12, s13, s14 and s15), and therefore they trade less intermediate inputs, show a lower Rsquared than some manufacturing sectors and construction (s6, s8 and s9 are always above the 80% Rsquared).

Hence, we added the following paragraph in p. 15 which includes new references from the relevant literature to support our argument:

> 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 and Kletzer (2005) challenge this conventional view of goods as tradable and services as non-tradable, in Gervais and Jensen (2019) analysis 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.

I find figure 4 a bit confusing. I would suggest to the author(s) to consider its replacement with a table of selected years given the overall trends are similar or the most interesting sectors. The full results could go to an appendix.

We do appreciate the above given that we have to plot 15 lines for 9 time periods. However, a table with, let's say, only 4 years still includes 4*15 = 60 cells and does not provide the bigger picture that Figure 4 offers. Therefore, we created an appendix, which, among other things includes a table with the R-squared values for the different sectors as presented in Figure 4.

Figure 5 is a bit messy

Figure 5 has now been amended.

- A local level application

I find this one of the main contributions of the paper. i.e. the ability to draw trade information for places where there is none. I suggest the author(s) can make more of it by discussing it more explicitly in the introduction but also other parts of the paper.

We have now added the following paragraph in the introduction in p. 2:

> 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.

In addition, we have created interactive visalusations of the hypelinks flows at the NUTS2 level and the trade flow predictions for LADs.

Reviewer: 3

The paper is well written and the research innovative and sound. A couple of main issues and a minor point could improve the quality of the manuscript.

Thank you very much for your nice comments.

First, there is no discussion or comparison justifying the need for ML approaches in general and for RF in particular. The author present very good results but do not discuss if such results are possible with other approaches. It would be great to see the results coming form the same approach but based on LASSO. Is an ML ensemble approach really needed? Would a statistical modeling approach(e.g., LASSO) result in data overfitting?

The current version of the paper now includes an estimation using LASSO and the same workflow. These results are inferior to the ones derived from RF and are briefly presented in the Appendix. We also included the following footnote in the second paragraph of Section 5:

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

Second, I found the discussion of variable selection too poor. First, there is no clear discussion, at least based on data. There are several approaches to extract relevant variables from the trees in the forest. I would suggest feature contribution measures such as the ones provided by the R package rfFC. Further the analysis provided by removing some variables is not fully discussed. The main results, by looking at the graphs and measures, are two and only one is discussed. the authors acknowledge the role of distance, and its temporal dynamics. But hyperlinks look like extremely irrelevant and for all periods. This point, which is one of the main assumptions of the paper is not discussed, but it must be.

Regarding the feature contribution and feature importance, our aim here is to build predictive models capable of making *out of sample* predictions for interregional trade flows. As you indicated our models are successful in doing that. We empirically tested the importance of the features we employed. Instead of presenting the results of the feature importance, which are based on in-sample estimations, we opted towards re-running our models and remove some of the features. We then tested these models in out of sample data and these results are presented in Figure 5 and discussed in second to last paragraph of section 5 (p. 15). This process allows us to assess the role of hyperlinks on out of sample interregional trade predictions. Indeed, the previous text was not entirely clear regarding the two new specifications, the results of which are presented in Figure 5. We have now corrected this and we explicitly highlight the value of our results in comparison to the alternative specifications presented here.

> 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.

We also tried to use the `rfFC` package as recommended, but unfortunately the functions from this package do not work anymore. We suspect that this is because the package is not maintained as its last update according to its github page is from 2015.

In addition, the paper now includes further justification of the other features included in the model. The below paragraph is placed in the Data section in p. 11:

> 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 et al., 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 regionalization 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.

Finally, I think the figures could be improved by being more specific on titles (with details of graph pane title when multiple panes are present) and axis titles. That information is always included in the caption, but would be easier for the reader to see it printed on the graph.

Figure 7 has now been amended.

Reviewer: 4

This is a relatively well organized and written manuscript that proposes interesting work to process digital linkages into origin-destination flows and use them as a proxy for interregional trade. The topic is interesting and well researched, and in particular, the authors provide a very rich background through the type of thorough literature review that is often missing from quantitative geographic analysis. I believe that in general there is a lot of merit and strong potential to use digital technologies to substitute more official data sets. And as these novel data sources are utilized and created, they also need to be validated and compared to those official data sets and other competitive sources. As it currently stands, this manuscript provides a quality contribution. Below are some more specific comments and suggestions that I think could improve the manuscript and further strengthen its contribution.

Thank so much for highlighting the value of and the quality of our contribution.

It wasn't clear to me what the material and material interdependencies referred to in the title. I think the second half of the title could be revised to better reflect the contribution of the manuscript. Perhaps something like, "A methodology for using the web to predict granular trade flows over time: data extraction, validation, and comparison".

We have now amended the paper title: Using the web to predict regional trade flows: data extraction, modelling, and validation. We hope this is clearer.

It is stated that, "Our underpinning hypothesis is that trade leaves behind digital breadcrumbs (Rabari and Storper 2014), which can be effectively utilised to predict interregional trade flows". I think it would be useful to revise this to PHYSICAL trade leaves behind digital breadcrumbs" it might be useful to even further clarify that these breadcrumbs can be used to reconstruct a proxy to those physical trade linkages, rather than predict them. I think the difference here is that one is a representation of another phenomenon, where as a prediction usually relates to some kind of model based on other phenomena.

We have now amended the above-mentioned sentence as per the below. We do understand the *proxy* point, but as we are not just building a proxy variable, so opted towards using the verb *capture* in the below.

> 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.

We have also added the word "physical" in the amended abstract.

I'm not sure that the paper utilizes the state of the art in machine learning. Random forests are now considered the bread and butter of any data science toolbox. It feels a little counter-intuitive to say that it is aligned with the state of the art in machine learning but to mention and rely on techniques that have essentially been around for about two decades and are now widely used. This is to say that the machine learning aspect of the paper is not really the primary or even the secondary contribution of this manuscript. I think it would be stronger to highlight those other things that are the primary and secondary contributions and then paragraphs, such as the first paragraph on page 3, and some of the text in the conclusion can be revised. There are certainly research threads out there that are actively looking at using more cutting edge machine learning techniques on origin destination type data sets. The fact that a more data driven technique is used here instead of more traditional parametric inferential statistical models is not new either. Neural networks, for example have been used spatial interaction models for decades, and there are examples of other machine learning techniques out there as well.

Point taken. What we wanted to emphasise was our aim to make out of sample predictions instead of estimating a traditional, explanatory, parametric model. Indeed, RF are well established estimators. We copy here the amended text from the Introduction:

> Our paper contributes to this line of inquiry by utilising novel web data and machine learning algorithms to make out-of-sample predictions for the UK NUTS2 regions during the period 2000-2010.
>
> 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.

and the Conclusions:

> Our paper aims to address this gap by proposing an innovative research framework, which is based on openly accessible web data and predictive models.

There is mention of policy implications in about the introduction and the conclusion, however I believe that this paper is largely methodological. The authors State,

"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"

And I believe that this is one of the largest implications of their work and a sufficient contribution, such that linking it to policy-making/implications can be relegated for future work.

We agree with the referee that the paper is essentially methodological, but at the same time another referee is asking for stating the policy implications and include more explicitly what we are missing by not having spatially detailed data. We hope the referee understands why we include in the paper the policy implications of our analysis both in the introduction and the conclusions sections. We believe that these discussions do not distract the readers from realising the value of our proposed research framework.

I think that it is useful to clarify throughout the manuscript that out of sample prediction here refers to the temporal dimension and not the spatial dimension. If I were modeling OD flows between pairs of locations, then I would think that out of sample refers to predicting the flows between an OD pair that was not included in the training data set. Since the model being used has no explicit time component, I think the correct terminology might be forecasting or extrapolation, but either way I think it would be useful to differentiate between spatial out of sample and temporal out of sample predictions. I'm also wondering about the density of the OD flows for each time. Are there any zero flowers in one year but not another year?

Thank you, this is useful. We opted towards using a temporal out of sample workflow instead of a spatial one, because this is more important and useful for the nature of our problem. Interregional trade data is missing for all pairs of OD, not only for a subset of them. We also use the term forecasting when we refer to our research framework: *rolling forecasting.* For the revised version of the paper we expanded the 'out of sample' term to 'temporal out of sample' throughout the manuscript. We copy below these instances.

> 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.
>
> 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).

> This workflow enables us to make temporal out-of-sample predictions and test our models and research framework in previously unseen data.
>
> Indeed, our models are able to make highly accurate temporal out of sample predictions for interregional trade in the UK.
>
> With the exception of these extreme values though (trade flows above £50
> billions) our model perform remarkably well in temporal out of sample predictions.
>
> 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.

Regarding zero flows, we do have OD pairs with zero hyperlinks for some years, but no pairs with zero trade flows. The number of pairs with zero hyperlinks decreases throughout the study period as expected.

It wasn't very clear to me the nature of the post codes that are used to Geolocate the digital linkages. Are these the postcards listed on each website? Can it be listed anywhere on a website? This is probably very straightforward if you have seen the data, but I think it could be useful to clarify this in the text.

We do appreciate that the unstructured nature of our data makes it not straightforward to communicate their specificities. The data section (4) of the paper included the following:

> 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.

In essence, the .html documents of all the archived .uk web pages were scanned to identify alphametric strings similar to UK postcodes. Our aggregation processes allows us to assess this information not at the webpage level, but instead at the website and, therefore, we use these terms carefully in the paper. We observe websites with a range of postcodes from one unique postcode to thousands as per Table 1. Our expectation is that the postcodes included in websites with only one unique postcode represent the trading addresses of these commercial activities. In other words, these are the 'contact us' type of webpages. We believe that this is a fair expectation as we know from the literature that commercial websites perform specific missions. See for instance the below amended text in p. 9:

> 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 et al., 2009: 364).

The following text is also included in Section 4, which further explains this process:

> we use this information to geolocated these wepbages and the websites these webpages are contained within.
>
> Firstly, we aggregate the Geoindex data from webpages to websites by grouping together all archived webpages which are contained under the same website[5].
>
> ---
>
> [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.

We firstly analyse websites with a unique postcode included in all their archived webpages. As per Blazquez and Domenech (2018) we expect these websites to represent economic activities trading in the unique location included in the archived webpages. Then, as a robustness check we repeat our analysis for websites with up to 10 unique postcodes and the results remain at large the same. Considering all the above, we are confident that our geolocation process is meaningful and informative.

The authors state that, "We opted against pooling the data to maintain their temporal structure both for methodological and conceptual reasons." Another potential reason is that if processes generating the data are changing over time, which seems likely for data aggregated yearly, then you would not want to train a model on data generated by a mix of processes, and less you were then going to try and protect data for a time frame that matches. More simply, I think a good reason for this strategy is to avoid overfitting as we don't reasonably suspect that the data generating processes stay the same over large spans of time. This has been demonstrated using spatial interaction models of various types, however I'm not sure if it has been done so in the context of trade flows yet (Oshan, 2020; Mozolin et al., 2000; Mikkonen & Luoma, 1999) .

Mozolin, M., Thill, J.-C., & Lynn Usery, E. (2000). Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. Transportation Research Part B: Methodological, 34(1), 53–73. https://doi.org/10.1016/S0191-2615(99)00014-4

Mikkonen, K., & Luoma, M. (1999). The parameters of the gravity model are changing–how and why? Journal of Transport Geography, 7(4), 277–283.

Oshan, T. M. (2020). Potential and Pitfalls of Big Transport Data for Spatial Interaction Models of Urban Mobility. The Professional Geographer, 72(4), 468–480. https://doi.org/10.1080/00330124.2020.1787180

Thank you. We certainly agree with this comment and we have added the following text a few lines below (p. 7)

> 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).

The figures could use some longer descriptions in their captions to make it clear what is being represented. For example, I am still not quite clear on the methodology: Digital trace networks are created for each of the years, and these are used to train models for two consecutive years and then predict the flows for the next year. Since we have data for all of the years, we can then validate those predictions. What could be clear to me is whether or not the explanatory variables in the model are also changing over time. Are these the same for each model that is trained? And do they change from the training step to the prediction step or is it only the dependent variable that changes. Demonstrating this latter scenario would be the strongest contribution for showing that if we trained a model, then we can use it to forecast flows before they would otherwise be available from any other data source, given that we can collect other types of data.

Yes, we matched yearly the explanatory variables with the outcome variable. We trained such biyearly models using data from years t and t + 1. Then, we use these trained models to make predictions for the next year t + 2. We do these predictions by using the explanatory variables for years t + 2. In other words, using the trained model and explanatory variables from years t + 2, we are able to predict interregional trade for for year t +2. We had tried to clarify this throughout the paper. For example, the Methodological Framework section includes the following text:

> To estimate RF models we employ the widely used caret package for R (Kuhn et al. 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.

To further clarify the process, we have added the following text in the Results section (p. 13):

> 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.

We hope that this addition helps. We opted against including this description from the Figures and their captions as we fear they will become too wordy. We changed through the title of Table 3:

> Accuracy metrics for predicting unseen data from t + 2

Finally, the manuscript could benefit from some additional grammatical editing and correction of typos. These were not typically problematic for understanding the content, but the further polishing would strengthen the contribution of this manuscript. on the JISC UK dataset and the construction of the extracted data instead of the importance of the framework.

The paper has been edited by a native English speaker.