ISSN 1610-241X (PRINT)
ISSN 1612-7501 (ONLINE)

# CESifo Economic Studies

Volume 64 Number 2 June 2018

http://academic.oup.com/cesifo

## On the Use of Geo-Coded Data in Economic Research





CESifo
A joint initiative of the University of Munich's Center for Economic Studies and the Ifo Institute

OXFORD UNIVERSITY PRESS

# CESifo Economic Studies

With the increasing specialisation in the profession, there is an urgent need for building bridges between different areas in economics. The economic research papers published in academic journals are becoming increasingly less accessible to those not working in the same field. *CESifo Economic Studies* bridges the gap between specialised research in economics and the interest of non-specialists. The journal combines theory and empirical research with a strong emphasis on policy issues. *CESifo Economic Studies* publishes contributed papers by top-level academic economists which are written in a style accessible to all economists regardless of their specialisation. The primary focus of the journal is on policy problems, but it also publishes high-quality, non-technical articles which are of general interest to economists.





Abstracted and Indexed in Social Sciences Citation Index, Social Scisearch, Current Contents/Social and Behavioural Sciences, EconLit, Journal of Economic Literature, Scopus and Geobase.

**SUBSCRIPTIONS**

A subscription to *CESifo Economic Studies* comprises 4 issues. Prices include postage by surface mail, or for subscribers in the USA and Canada by airfreight, or in India, Japan, Australia and New Zealand, by Air Speeded Post. Airmail rates are available on request.

Annual Subscription Rate (Volume 64, 4 issues, 2018)
*Institutional*
Print edition and site-wide online access: £314/€465/US$601
Print edition only: £289/€429/US$553
Site-wide online access only: £246/€365/US$471
*Personal*
Print edition and individual online access: £68/€101/US$130
Please note: US$ rate applies to US & Canada, Euros applies to Europe, UK£ applies to UK and Rest of World

There are other subscription rates available, for a complete listing please visit http://academic.oup.com/cesifo/subscriptions.

Full prepayment, in the correct currency, is required for all orders. Orders are regarded as firm and payments are not refundable. Subscriptions are accepted and entered on a complete volume basis. Claims cannot be considered more than FOUR months after publication or date of order, whichever is later. All subscriptions in Canada are subject to GST. Subscriptions in the EU may be subject to European VAT. If registered, please supply details to avoid unnecessary charges. For subscriptions that include online versions, a proportion of the subscription price may be subject to UK VAT. Personal rate subscriptions are only available if payment is made by personal cheque or credit card and delivery is to a private address. For further details, see inside back cover.

The current year and two previous years' issues are available from Oxford Journals. Previous volumes can be obtained from the Periodicals Service Company at http://www.periodicals.com/oxford.html or Periodicals Service Company, 11 Main Street, Germantown, NY 12526, USA. E-mail: psc@periodicals.com. Tel: (518) 537-4700. Fax: (518) 537-5899.

*CESifo Economic Studies* (ISSN 1610-241X) is published four times per year in March, June, September and December by Oxford University Press, Oxford, UK and distributed in the USA by Central Mailing Services c/o UKP Worldwide, 1637 Stelton Road B2, Piscataway, NJ 08854.

The US annual print subscription price is $601.00. Airfreight and mailing in the USA by agent named Central Mailing Services c/o UKP Worldwide, 1637 Stelton Road B1-2, Piscataway, NJ 08854. Periodicals postage paid at Piscataway, NJ and additional mailing offices.

US Postmaster: send address changes to Central Mailing Services c/o UKP Worldwide, 1637 Stelton Road B1-2, Piscataway, NJ 08854.



Oxford Journals Environmental and Ethical Policies
Oxford Journals is committed to working with the global community to bring the highest quality research to the widest possible audience. Oxford Journals will protect the environment by implementing environmentally friendly policies and practices wherever possible. Please see https://academic.oup.com/journals/pages/about_us/ethical_policies for further information on Oxford Journals' environmental and ethical policies.

# CESifo Economic Studies

Volume 64 Number 2 June 2018

http://academic.oup.com/cesifo

## Contents

### On the Use of Geo-Coded Data in Economic Research*



*The Symposium papers have been selected from those presented at a CESifo Economic Studies conference on the use of geo-coded data in economic research in November 2016, organised by Ina Blind, Matz Dahlberg and Gustav Engström. Conference organisers were the guest editors of this Symposium.

*CESifo Economic Studies*, 2018, 127–149
doi: 10.1093/cesifo/ify004
Advance Access Publication Date: 16 March 2018
Original article

# Geographic Natural Experiments with Interference: The Effect of All-Mail Voting on Turnout in Colorado

**Luke Keele* and Rocío Titiunik†**

*McCourt School of Public Policy and Government Department, Georgetown University, 304 Old North, 37th & O St, Washington, DC, 20057, USA. e-mail: lk681@georgetown.edu and †Department of Political Science, University of Michigan, 5700 Haven Hall, 505 South State St, Ann Arbor, MI 48109, USA. e-mail: titiunik@umich.edu

## Abstract

We analyze a geographic natural experiment during the 2010 Colorado primary election in the USA, when counties in the state of Colorado had the option to have an all-mail election or retain traditional in-person voting on Election Day. The town of Basalt, in the southwestern part of the state, is split in half by two counties that chose different modes of voting. Our research design compares these two counties to understand whether turnout levels were altered by all-mail elections. Our analysis considers the possibility that social interactions may lead to spillover effects—a situation in which one unit's outcome may be affected by the treatment received by other units. In our application, treated and control voters lived in very close proximity and spillovers are possible. Using the potential outcomes framework, we consider different estimands under the assumption that interference occurs only when treated individuals are in close geographic proximity to a sufficiently high number of control individuals. Under our assumptions, our empirical analysis suggests that all-mail voting decreased turnout in Colorado, and shows no evidence of spatial interference between voters. (JEL codes: C18, C99)

**Key words**: econometric and statistical methods, spatial models

## 1. Introduction

In recent decades, the USA has witnessed an increase in methods of voting that differ from traditional in-person voting on Election Day. These reforms, commonly referred to as 'convenience voting', include in-person early voting (where voters may cast a vote in person before election day), no-excuse absentee voting (where voters may apply for an absentee ballot without providing a reason for doing so), and all-mail voting (where voting by mail is mandatory)—see Gronke et al. (2008) for a review. These policies are often implemented with the goal of reducing the costs of voting, which is in turn expected to increase voter participation.

Among the different convenience voting policies that have been adopted, all-mail voting is the most drastic, since, under this policy, in-person precinct voting is eliminated and there are no polling places; instead, citizens receive a ballot in the mail several weeks in advance of Election Day and then return it by mail to the election administration office. Since all-mail voting is the only convenience method that eliminates precinct-place voting, its effects on turnout could be different from the effects of other types of convenience voting. While all-mail voting may be more convenient, the move to all-mail elections reduces the social aspect of voting, which can be a key motivator for political participation (Gerber et al. 2008). All-mail voting also removes the possibility of using Election Day as a focal point for mobilization efforts by political parties. Combined, these factors might result in fewer voters casting a vote.

Given the far-reaching nature of all-mail voting reforms, scholars have been interested in studying whether they affect voter turnout. As various states have implemented vote-by-mail systems either on a statewide or more local basis, a number of studies have attempted to estimate whether this mode of voting increases or decreases turnout. Much of the focus has been on the state of Oregon, where polling-place voting was gradually eliminated during the 1990s, and since 1998 all statewide primary and general elections are conducted by mail only. Several studies have concluded that Oregon's all-mail voting reform increased turnout, though the estimated magnitude of the change varies considerably from study to study (Southwell and Burchett 2000; Karp and Banducci 2000; Berinsky et al. 2001; Richey 2008; Gronke and Miller 2012). Some additional evidence on this question comes from other states. Kousser and Mullin (2007) and Bergman and Yates (2011) study California, where county election officials can assign voters to all-mail voting precincts in low-population areas, and find that turnout appears to be lower under an all-mail voting system. Gerber et al. (2013) study the large-scale move from polling-place to all-mail elections in the state of Washington, and they find that all-mail voting increases turnout by 2–4 percentage points.

We examine this question using a geographic natural experiment in Colorado, where in 2010 counties were given the choice to require that votes be cast by mail during the primary election. Counties that adopted all-mail elections removed other alternative methods of voting, while counties that did not adopt all-mail voting still offered traditional polling-place voting on Election Day (and also allowed by-mail no-excuse absentee voting). In general, given that voter administration is conducted by county governments, counties may chose their mode of voting to try to accomplish their specific voter turnout goals. This type of strategic decision-making may complicate naive statistical inferences that simply compare all-mail counties to in-person counties. In an attempt to minimize these complications, our study focuses on voters in Basalt, a town that is split by the border between Eagle county, which adopted all-mail election voting, and Pitkin county, which retained in-person voting. Our research design focuses on a narrow area around the boundary between both counties, and makes the assumption that, within this small area, voters in the town of Basalt are split in a haphazard fashion between Eagle and Pitkin counties, after conditioning on covariates. Based on this research strategy, we draw inferences about the effects of all-mail elections on voter turnout.

A key element of our research design is our focus on a small geographic area around the boundary that separates both counties, since citizens who reside close to the county border on either side share important predetermined characteristics that may be related to voter turnout decisions. This focus on treated and control voters who reside near each other, while increasing comparability, may also increase the likelihood that treated voters interact with control voters in a way that affects their outcomes. The presence of interference between voters would undermine the interpretation of our estimates as the average treatment effect of all-mail voting. In general, research designs based on comparisons of units who reside in close geographic proximity to each other are more likely to suffer from spatial forms of interference between units, and will face a trade-off between increasing comparability and reducing interference (Keele et al. 2017). However, the patterns of interference between units may be more general than those induced by residential proximity. For example, workers could be influenced by colleagues whose residence is geographically far from their own, but with whom they interact at the workplace. Even more drastically, interference could arise between units who are never geographically or physically close in any capacity, in particular via social media interactions. The framework we use below could be applied to these more general forms of interference, replacing our notion of residential proximity with a notion of proximity in a social network, assuming that the latter is known—for an example of a study that considers both types of proximity, see Verdery et al. (2012).

One goal in our study is to understand whether interference alters our inferences about the effect of all-mail voting. Our empirical analysis builds on prior work that employs the potential outcomes framework to study interference between units. Sobel (2006) characterizes several estimands of interest under interference, and shows that the usual difference-in-means estimator in a completely randomized experiment is no longer unbiased for the average treatment effect. Hudgens and Halloran (2008) consider a two-stage randomization, in which interference occurs within but not between groups, and define direct and indirect causal effects that consider how the outcomes of one unit change as the treatment assignment of all other units stays constant or changes. Hierarchical models in which interference occurs within but not between groups are also considered, among others, by Tchetgen Tchetgen and VanderWeele (2012), Vanderweele (2008), and VanderWeele et al. (2013). Aronow and Samii (2017) consider the estimation of average causal effects under general forms of known interference. Rosenbaum (2007) and Bowers et al. (2013) consider hypothesis testing under interference in a Fisherian framework. Gerber and Green (2012) consider the problem of spatial interference in randomized experiments, and Sinclair et al. (2012) design a multilevel voter-mobilization experiment to detect spillovers within and between households.

Methodologically, our approach is most similar to the setting in Hong and Raudenbush (2006) and Verbitsky-Savitz and Raudenbush (2012). Hong and Raudenbush (2006) study the effect of retaining low-achieving children in kindergarten versus promoting them to 1st grade, and model interference effects by means of a scalar function of the treatment assignment vector within each school. Verbitsky-Savitz and Raudenbush (2012) apply the ideas in Hong and Raudenbush (2006) to a spatial setting to study the effect of a community policing program on neighborhoods' crime rates in Chicago. They assume that the potential outcome of a given unit depends on the other units' potential outcomes via a scalar function that contains the proportion of contiguous units.

In our analysis of the effects of all-mail voting under spatial interference, we also impose the assumption that interference is a scalar function of the treatment assignment vector. In particular, we assume that interference is a function of geographic proximity to a sufficiently dense area of voters of the opposite treatment status, which vastly reduces the number of potential outcomes for every unit and leads to two estimands of interest.

This approach is similar to the approach in Verbitsky-Savitz and Raudenbush (2012), who model interference as a function of the proportion of contiguous units, although we impose an additional restriction. Our setup treats the geographic locations of the voters in our sample as random, and the boundary between treated and control areas as fixed—an approach that is particularly well suited to geographical natural experiments that focus on a narrow band around a boundary, and differs from other approaches that allow for spatial interference under the assumption that geographic locations are given. Our function of interference can be modified to reflect particular patterns of geographic spillovers.

The remainder of the article is organized as follows. In Section 2 we describe the Colorado application in more detail. In Section 3 we present our notation and describe our causal estimand under the assumption of no interference. In Section 4 we allow for geographic interference, and explore some further issues in Section 5. In Section 6, we use this framework to estimate the treatment effect of Colorado's all-mail voting on turnout, first ignoring interference between units, and then allowing for interference based on residential proximity (Section 6.1). In this section, we also explore sensitivity to differential registration (Section 6.2). We offer concluding remarks in Section 7.

## 2. All-Mail Voting in 2010 Colorado Primary

In recent years, the state of Colorado has implemented several reforms aimed at making voting more convenient. Starting in 2008, voters could choose to be placed on a permanent vote-by-mail list. For the 2010 primary, the Secretary of State allowed each county to choose whether to hold either all-mail elections, use voter centers, or hold traditional in-person voting at precincts. Figure 1 contains a map showing the mode of election selected by each county. While urban areas generally selected all-mail elections, many rural counties chose to use in-person voting.

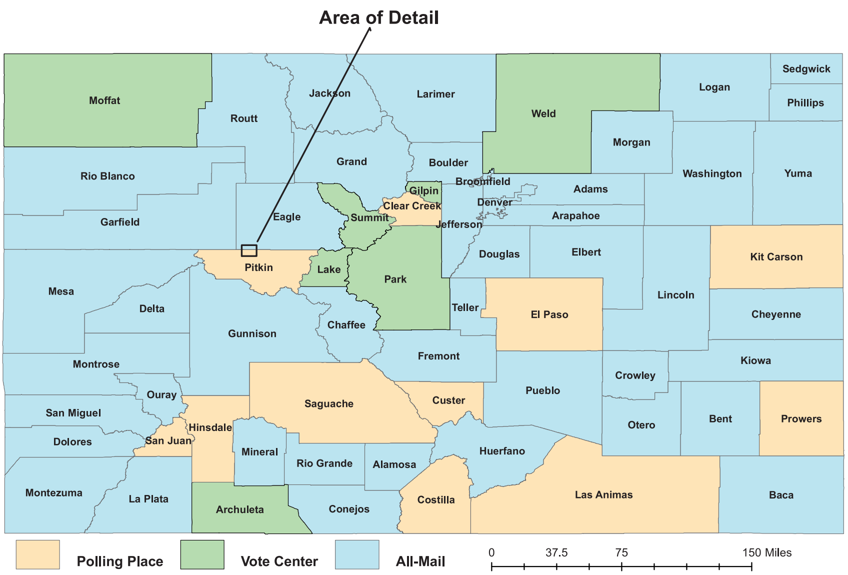


**Figure 1.** County map of Colorado with model of voting and location of Basalt highlighted.

The different modes of voting chosen by different counties allow us to study how election mode affects voter turnout. However, given that counties are responsible for election administration and were able to select their preferred mode of election in the 2010 primary, comparisons across counties might be invalidated by unobserved confounding or heterogeneity. To minimize this concern, we looked for some location where a town or city is split by a county border, where one county uses all-mail voting, while the other county uses in-person voting. We found that one town in the southwestern part of the state, Basalt, was split exactly in this fashion.

Figure 1 shows the location of the town of Basalt. According to the 2010 census, Basalt has a total population of 3857. The population is largely White, and about 20% of it identifies as Hispanic. The town is close to the resort city of Aspen, and using property sale records, we found that the median house price in 2010 was over $600,000. Figure 2 contains a map showing the town in greater detail. The central part of Basalt is split by the county border which defines mode of election, and this part of the town contains the main shopping district, residential areas, and schools. While the county border splits the town, the entire area is within the same school district. Moreover, all residents of Basalt attend the same set of public schools which are located within the central part of the town. Although property taxes in Colorado have a county component, property taxes are based on five different tax zones with school district contributing the most to the overall property tax burden.

Primary elections often hold little interest for voters, since primary races are often uncompetitive. The 2010 Colorado primary, however, had three high-profile elections on the ballot. In the Republican gubernatorial primary, a Tea Party insurgent beat Scott McInnis, a six-term U.S. representative, after it came to light that McInnis plagiarized a water study

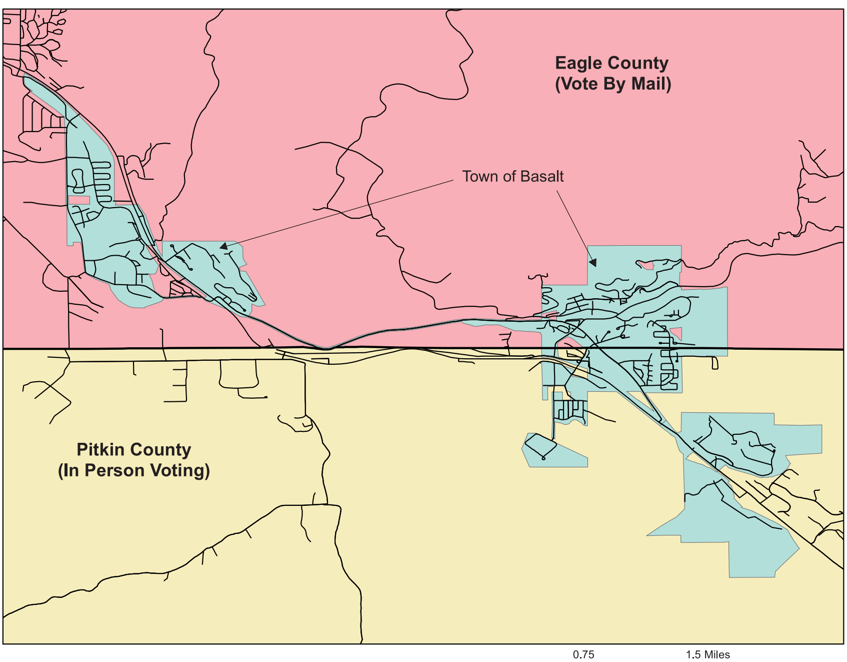


**Figure 2.** The county discontinuity: Basalt split by county boundary.

he was paid to conduct. In the Democratic US Senate primary, the candidate endorsed by then-President Obama narrowly beat a more liberal candidate endorsed by former president Bill Clinton. In the Republican US primary Ken Buck, a Tea Party candidate, beat Jane Norton—the candidate endorsed by the Colorado Republican party establishment. Results from the primary received national coverage and were featured on the front page of the *New York Times*.

## 3. Estimation and Inference without Interference

We adopt the potential outcomes framework for causal inference (see Holland 1986; Rubin 2005) assuming first that voters do not interfere with one another. We generalize the framework to allow for interference in Section 4.

We conduct our analysis with individual-level voter data. We use the binary variable $D_i \in [0, 1]$ to denote treatment status for resident $i$, with $D_i = 1$ if $i$ resides in Eagle County and is assigned to an all-mail election, and $D_i = 0$ if $i$ resides in Pitkin county and may vote in-person on Election Day. Each resident has several potential outcomes, only one of which is realized by the assignment of treatment. There are also $k$ predetermined covariates for each resident, which we denote by $\boldsymbol{X}_i$. The observed data are $\{Y_i, D_i, \boldsymbol{X}_i\}_{i=1}^n$, which we assume is an i.i.d. random sample from a larger population. We collect all treatment indicators in the $n$-vector $\boldsymbol{D}$, and let $Y_i(\boldsymbol{D})$ be the potential outcome of resident $i$. We denote the observed outcome by $Y_i \equiv Y_i(\boldsymbol{d})$, where $\boldsymbol{d}$ is the realized treatment assignment vector. In general, if we let the treatment status of every resident affect the potential outcome of every other resident, every $i$ will have one distinct potential outcome for every value that the treatment vector $\boldsymbol{D}$ might take, which is $2^n$. We start by entirely simplifying this problem and assuming that there is no interference between residents, as formalized in the following assumption.

**Assumption 1** (No Interference). The potential outcome of each unit depends only on the treatment received by that unit and not on the treatment assigned to any other unit: for all $\dot{\boldsymbol{D}} \neq \ddot{\boldsymbol{D}},\ Y_i(\dot{\boldsymbol{D}}) = Y_i(\ddot{\boldsymbol{D}})$ if $\dot{D}_i = \ddot{D}_i$, for $i = 1, 2, \ldots, n$.

Under Assumption 1, we can write $Y_i(\boldsymbol{D}) = Y_i(D_i)$, since $i$'s potential outcome only depends on the treatment received by $i$. In this case, the observed outcome simplifies to $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$. The quantity $\tau_i = Y_i(1) - Y_i(0)$ captures the effect of all-mail voting for the $i$th voter. Our interest is estimating the average treatment effect on the treated (ATT), $\tau = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$, which is only defined under the assumption of no interference.

### 3.1 A geographic identification strategy

We estimate the effects of all-mail voting on voter turnout using a geographic natural experiment. Under this identification strategy, a geographic or administrative boundary splits units into two adjacent areas, one of which receives a treatment, $\mathcal{A}^t$, and the other of which receives control, $\mathcal{A}^c$, and analysts make the case that the assignment of units into treated and control areas occurs in an as-if random fashion (Keele and Titiunik 2015, 2016). Researchers make comparisons between units in the treated and control areas to infer the effect of the treatment on an outcome of interest, relying on the spatial proximity of each unit to the border between $\mathcal{A}^c$ and $\mathcal{A}^t$, and on the fact that the treatment changes abruptly along this boundary—that is, $D_i = 1$ if unit $i$ is located in $\mathcal{A}^t$, and $D_i = 0$ if $i$ is located in $\mathcal{A}^c$. Applying this strategy to our application, we assume that around the county border that divides the town of Basalt into all-mail and in-person voting regimes, individuals choose their residence on either side of the county boundary on an as-if random fashion, possibly after conditioning on predetermined covariates.

In essence, the assumptions behind a geographic natural experiment require that the placement of each unit on either side of the geographic boundary between $\mathcal{A}^c$ and $\mathcal{A}^t$ be as-if random or, in other words, that units cannot precisely sort or self-select to one side of the boundary based on unobserved factors that are also correlated with the outcomes of interest. A consequence of this assumption is that observable predetermined covariates should be similar in expectation within some narrow area around the border of interest. A weaker assumption is that treatment assignment is as-if randomized for those who live near the border, after conditioning on a set of observable covariates (Keele et al. 2015). Given the need to condition on covariates, such designs have been characterized as geographic-quasi experiments (GQEs) (Galiani et al. 2017; Keele et al. 2017 ). Since we are interested in the ATT, we adopt a version of this assumption that only restricts the average potential outcome under control: we assume that there exists a small neighborhood around the boundary that separates both areas where the average potential outcome under control is mean independent of the treatment given the covariates. We state it formally below.

**Assumption 2** (Conditional Mean Independence in Local Neighborhood). For all units that reside in a narrow band around the boundary that separates $\mathcal{A}^c$ and $\mathcal{A}^t$, $\mathbb{E}[Y_i(0)|\boldsymbol{X}_i, D_i = 1] = \mathbb{E}[Y_i(0)|\boldsymbol{X}_i, D_i = 0]$.

Note that Assumption 2 implicitly invokes Assumption 1, which reduces the set of potential outcomes to $Y_i(1)$ and $Y_i(0)$. Moreover, since our boundary of interest is simultaneously the boundary of multiple institutional, administrative, or political units, and we wish to make inferences about the effect of only one of these treatments, we must assume that the treatment of interest is the only treatment that affects potential outcomes, i.e. that there are no compound treatments (Keele and Titiunik 2015). In particular, in our application we must assume that no other county-level factor affects voter turnout other than the administration of voting.

Under these assumptions, the ATT is identified by

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \\ &= \mathbb{E}[\tau(\boldsymbol{X}_i)|D_i = 1] \\ &= \mathbb{E}[\mathbb{E}(Y_i|\boldsymbol{X}_i, D_i = 1) - \mathbb{E}(Y_i|\boldsymbol{X}_i, D_i = 0)|D_i = 1],\end{aligned}$$

where we have defined $\tau(\boldsymbol{X}_i) \equiv \mathbb{E}(Y_i(1) - Y_i(0)|D_i = 1, \boldsymbol{X}_i)$, the ATT conditional on covariates.

We estimate effects of interest using least squares methods. To motivate our estimation strategy, we first assume that treatments were randomly assigned. We can express the observed outcome as:

$$Y_i = \mu + \tau \cdot D_i + u_i, \tag{1}$$

where

$$\mu = \mathbb{E}[Y_i(0)|D_i = 0]$$
$$u_i = (1 - D_i) \cdot \{Y_i(0) - \mathbb{E}[Y_i(0)|D_i = 0]\} + D_i \cdot \{Y_i(1) - \mathbb{E}[Y_i(1)|D_i = 1]\},$$

and $\tau$ is the ATT defined above. Random assignment would imply $\mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)|D_i = 1]$, which in turn would lead to $\mathbb{E}[u_i \cdot D_i] = \mathbb{E}[u_i|D_i = 1] = 0$. Thus, under random assignment, the coefficients $\mu$ and $\tau$ can be consistently estimated using least squares methods, simply regressing the voter turnout binary outcome on an indicator variable for treatment.

In our application, however, we do not believe that the condition $\mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i(0)|D_i = 1]$ is plausible. Instead, we assume Assumption 2 holds, where mean independence holds conditional on $\boldsymbol{X}_i$. For simplicity, to be able to apply least-squares methods to this case, we assume that we can condition on covariates in a linear fashion. In this case, we can express the observed outcome as:

$$\begin{aligned} Y_i &= \mathbb{E}[Y_i(0)|D_i = 0, \boldsymbol{X}_i] + \tau \cdot D_i + \tilde{u}_i \\ &= \boldsymbol{X}_i{}'\beta + \tau \cdot D_i + \tilde{u}_i \end{aligned}, \tag{2}$$

where

$$\begin{aligned} \tilde{u}_i = {} & D_i \cdot (\tau(\boldsymbol{X}_i) - \tau) + (1 - D_i) \cdot (Y_i(0) - \mathbb{E}[Y_i(0)|D_i = 0, \boldsymbol{X}_i]) \\ & + D_i \cdot (Y_i(1) - \mathbb{E}[Y_i(1)|D_i = 1, \boldsymbol{X}_i]), \end{aligned}$$

and we imposed $\mathbb{E}[Y_i(0)|D_i = 0, \boldsymbol{X}_i] = \boldsymbol{X}_i{}'\beta$ to obtain the second line. Now, under Assumption 2, $\mathbb{E}[\tilde{u}_i|D_i = 0, \boldsymbol{X}_i] = \mathbb{E}[\tilde{u}_i|\boldsymbol{X}_i, D_i = 1] = 0$, so $\tau$ can again be consistently estimated with least squares methods.

## 4. Interference between Units in a GQE

In geographical natural experiments generally, and the GQE in particular, the research strategy is based on a comparison of units that are spatially proximate, under the assumption that units very close to one another but with opposite treatment status can provide valid counterfactuals for each other. However, in some cases, this focus on spatially proximate units may introduce the possibility of spillovers or interference between units. In our application, we are concerned about indviduals who reside in the all-mail-voting (treated) area being influenced by individuals who reside in the in-person-voting (control) area because in the latter area Election Day still acts as a focal point, and the act of voting is socially coordinated and shared. This might make the election generally more salient in the in-person area in the weeks before the election, affecting the propensity to vote of all-mail residents on the other side of the boundary. We assume that interference between units depends on geographic proximity to units of opposite treatment status. This is similar to the approach in Verbitsky-Savitz and Raudenbush (2012), although we do not focus on contiguity but rather on the density of control units that reside within a pre-specified distance of a given treated unit. Since we are not similarly concerned about individuals in the in-person-voting area being affected by individuals in the all-mail-voting area, we assume that interference is one-sided, from control to treated—but the setup is generalizable to two-sided patterns of interference.

Thus far, we assumed that resident $i$'s treatment status did not depend on the treatment status of any other resident, allowing us to write potential outcomes as $Y_i(\boldsymbol{D}) = Y_i(D_i)$, since $i$'s potential outcome only depended on $i$'s own treatment status. If we allow for any pattern of interference, we must work with the full vector $Y_i(\boldsymbol{D})$, which allows individual $i$'s treatment status to depend on the treatment status of every other individual. However, this level of generality is unworkable, as the large number of causal effects per subject makes it difficult to summarize the data in any interpretable way.

To add structure and reduce the dimensionality of our problem, we assume that each individual's potential outcome depends on the individual's own treatment status, $D_i$, and also on the number of individuals of the opposite treatment status who reside within a specified distance of $i$'s location. To introduce the necessary notation, we first define the function $g_i(\boldsymbol{D}; \delta, \eta)$ for fixed values of the scalars $\delta$ and $\eta$, as follows:

$$g_i(\boldsymbol{D}; \delta, \eta) = \mathbb{1}(D_i = 1) \cdot \mathbb{1}(N_{i\delta} \geq \eta),$$

where

$$N_{i\delta} = \sum_{j=1}^{n} \mathbb{1}(d(i,j) \leq \delta) \cdot \left(D_j \neq D_i\right),$$

$\mathbb{1}(\cdot)$ is the indicator function, $d(i, j)$ is a measure of distance between $i$'s and $j$'s locations and $\delta, \eta \in \mathbb{R}$.

The function $g_i(\cdot)$ is an indicator for whether individual $i$ receives interference, taking a value of 1 or 0 for every individual. If individual $i$ is treated (i.e., resides in the all-mail-voting area), $g_i(\boldsymbol{D}; \delta, \eta) = 1$ if there are at least $\eta$ control individuals who reside within $\delta$ meters of $i$'s location, and $g_i(\boldsymbol{D}; \delta, \eta) = 0$ if there are less than $\eta$ control individuals in a $\delta$ radius around $i$'s location. If individual $i$ is control (resides in the in-person-voting area), $g_i(\boldsymbol{D}; \delta, \eta) = 0$ regardless of how many treated individuals reside close to $i$—because we have assumed one-sided interference only.

Thus, we capture our model of one-sided geographic-based interference by letting the potential outcomes depend on the full vector of treatment assigments $\boldsymbol{D}$ in a restricted way; in particular, we let individual $i$'s potential outcome depend on its own treatment status $D_i$ and on the treatment status of other individuals only through the function $g_i(D; \delta, \eta)$. Thus, in our interference framework, $Y_i(\boldsymbol{D}) = Y_{i;\delta,\eta}(D_i, g_i(\boldsymbol{D}; \delta, \eta))$, reducing the number of arguments in each individual's potential outcome from $n$ to 2. This is similar to the approach in Hong and Raudenbush (2006) and Verbitsky-Savitz and Raudenbush (2012), where the dependence of $i$'s potential outcomes on the treatment assignment of all other units is also modeled via a scalar function that substantially reduces the range of possible potential outcomes that may occur.

Under our specific assumption of geographic-based interference, every unit has three potential outcomes:

- $Y_{i;\delta\eta}(1, 1)$: Individual $i$ resides in the all-mail-voting area and receives spillovers, i.e. there are at least $\eta$ control individuals in the in-person-voting area within $\delta$ meters of $i$'s location.

- $Y_{i;\delta\eta}(1,0)$: Individual $i$ resides in the all-mail-voting area and does not receive spillovers, i.e. there are less than $\eta$ control individuals in the in-person-voting area within $\delta$ meters of $i$'s location.
- $Y_{i;\delta\eta}(0,0)$: Individual $i$ resides in the in-person-voting area. Since $g_i(\boldsymbol{D};\delta,\eta)=0$ for all individuals in the in-person-voting area, we can simply write $Y_{i;\delta\eta}(0,0)=Y_i(0)$.

We define the ATT in the absence of interference,

$$\tau_{\mathrm{T},\delta\eta}=\mathbb{E}\left[Y_{i;\delta\eta}(1,0)-Y_i(0)|D_i=1\right].$$

For brevity, we refer to this effect as the 'interference-free treatment effect'. In the context of our application, $\tau_{\mathrm{T},\delta\eta}$ captures the average effect of all-mail voting when individuals in the all-mail-voting area are geographically far from densely populated areas in the in-person-voting area and thus receive no spillovers. Since, under our model of interference, residents in the all-mail-voting area receive no spillovers when they are geographically distant from dense control areas, and there are no spillovers for control individuals, the two potential outcomes in $\tau_{\mathrm{T},\delta\eta}$ reflect the potential outcomes that would be observed under treatment and control in the absence of interference.

We also define an additional parameter,

$$\tau_{\mathrm{S},\delta\eta}=\mathbb{E}\left[Y_{i;\delta\eta}(1,1)-Y_{i;\delta\eta}(1,0)|D_i=1\right],$$

which captures the average effect of interference or spillovers on treated units. For brevity, we refer to this parameter as the 'interference effect'. In our application, $\tau_{\mathrm{S},\delta\eta}$ compares the average potential outcomes under all-mail voting for residents in the all-mail-voting area who are geographically close to the (densely populated parts of the) in-person-voting area, to the average potential outcome under all-mail voting for residents in the all-mail-voting area who are relatively isolated from the in-person area. Note that, in the absence of interference, $Y_{i;\delta\eta}(1,1)=Y_{i;\delta\eta}(1,0)=Y_i(1)$ for all $i$, which implies $\tau_{\mathrm{S},\delta\eta}=0$. Thus, a test of interference can be based on a test of the null hypothesis $H_0:\tau_{\mathrm{S},\delta\eta}=0$.

Both $\tau_{\mathrm{T},\delta\eta}$ and $\tau_{\mathrm{S},\delta\eta}$, however, depend on three different potential outcomes, and only one of those is observed for every $i$. We now investigate assumptions that, in the particular context of geographic natural or quasi experiments, could be invoked to identify these parameters. We first note that, letting $Y_i$ denote the observed outcome for individual $i$, we have the following equalities between observed and potential outcomes:

- $Y_i=Y_i(0)$ if $D_i=0$
- $Y_i=Y_{i;\delta\eta}(1,1)$ if $D_i=1$ and $g_i(\boldsymbol{D};\delta,\eta)=1$
- $Y_i=Y_{i;\delta\eta}(1,0)$ if $D_i=1$ and $g_i(\boldsymbol{D};\delta,\eta)=0$.

We now consider the following assumption,

**Assumption 3** (As-if random geographic location within interference areas).

$$\mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=1,\boldsymbol{X}\right]=\mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=0,\boldsymbol{X}\right]$$
$$\mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=1,\boldsymbol{X}\right]=\mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=0,\boldsymbol{X}\right]$$

A sufficient condition for Assumption 3 is that each unit is randomly assigned to a geographic location in the combined treated area, so that whether they fall in the interference region ($g_i(\boldsymbol{D};\delta,\eta)=1$) or the non-interference region ($g_i(\boldsymbol{D};\delta,\eta)=0$) is unrelated to their potential outcomes. Our assumption is weaker than this, since it requires only that conditional on pretreatment covariates, falling in the interference region in the treatment area is mean independent of potential outcomes, though it still is a strong assumption. However, as we discuss in detail below, under the type of treatment assignment that is typical of geographic natural or quasi experiments, this assumption might be plausible if the neighborhood around the boundary that separates treated and control areas is small enough and enough pretreatment covariates are available.

Under Assumptions 2 and 3, we have:

$$\begin{aligned}\tau_{\mathrm{T},\delta\eta}(\boldsymbol{X}) &\equiv \mathbb{E}\left[Y_{i;\delta\eta}(1,0)-Y_i(0)|D_i=1,\boldsymbol{X}\right]\\ &= \mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=0,\boldsymbol{X}\right]-\mathbb{E}[Y_i(0)|D_i=0,\boldsymbol{X}]\\ &= \mathbb{E}[Y_i|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=0,\boldsymbol{X}]-\mathbb{E}[Y_i|D_i=0,\boldsymbol{X}]\end{aligned}$$

and

$$\begin{aligned}\tau_{\mathrm{S},\delta\eta}(\boldsymbol{X}) &\equiv \mathbb{E}\left[Y_{i;\delta\eta}(1,1)-Y_{i;\delta\eta}(1,0)|D_i=1,\boldsymbol{X}\right]\\ &= \mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=1,\boldsymbol{X}\right]-\mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=0,\boldsymbol{X}\right]\\ &= \mathbb{E}[Y_i|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=1,\boldsymbol{X}]-\mathbb{E}[Y_i|D_i=1,g_i(\boldsymbol{D};\delta,\eta)=0,\boldsymbol{X}].\end{aligned}$$

These results, which express $\tau_{\mathrm{T},\delta\eta}(\boldsymbol{X})$ and $\tau_{\mathrm{S},\delta\eta}(\boldsymbol{X})$ exclusively in terms of observable data, allow us to estimate and make inferences about the treatment effect in the absence of interference, $\tau_{\mathrm{T},\delta\eta}$, and the interference effect, $\tau_{\mathrm{S},\delta\eta}$.

As above, we outline an estimation strategy using least squares methods. To simplify the notation, let $G_i\equiv g_i(\boldsymbol{D};\delta,\eta)$. Under Assumptions 2 and 3, we can express the observed outcome as:

$$\begin{aligned}Y_i &= \mathbb{E}[Y_i(0)|D_i=0,\boldsymbol{X}_i]+\tau_{\mathrm{T},\delta\eta}\cdot D_i+\tau_{\mathrm{T},\delta\eta}\cdot G_i\cdot D_i+\tilde{\epsilon}_i\\ &= \boldsymbol{X}_i{}'\gamma+\tau_{\mathrm{T},\delta\eta}\cdot D_i+\tau_{\mathrm{S},\delta\eta}\cdot G_i\cdot D_i+\tilde{\epsilon}_i\end{aligned}, \tag{3}$$

where

$$\begin{aligned}\tilde{\epsilon}_i = {}&(Y_i(0)-\mathbb{E}[Y_i(0)|D_i=0,\boldsymbol{X}_i])\cdot(1-D_i)+\\ &(Y_i(1,0)-\mathbb{E}[Y_i(1,0)|D_i=1,G_i=0,\boldsymbol{X}_i])\cdot D_i(1-G_i)+\\ &(Y_i(1,1)-\mathbb{E}[Y_i(1,1)|D_i=1,G_i=1,\boldsymbol{X}_i])\cdot D_iG_i+\\ &D_i\left(\tau_{\mathrm{T},\delta\eta}(\boldsymbol{X}_i)-\tau_{\mathrm{T},\delta\eta}\right)+D_i\cdot G_i\cdot\left(\tau_{\mathrm{S},\delta\eta}(\boldsymbol{X}_i)-\tau_{\mathrm{S},\delta\eta}\right)\end{aligned},$$

and we again impose the linear specification $\tilde{\alpha}\equiv\mathbb{E}[Y_i(0)|D_i=0,\boldsymbol{X}_i]=\boldsymbol{X}_i{}'\gamma$. Given our assumptions, $\mathbb{E}[\tilde{\epsilon}_i|D_i,G_i,\boldsymbol{X}_i]=0$, and the parameters can be consistently estimated by least-squares. We note, however, that our identification assumptions do not rely on linearity assumptions and other, more flexible estimators could be employed to estimate the parameters of interest.

### 4.1 The treatment assignment mechanism in geographic natural and quasi experiments

We now discuss the plausibility of Assumptions 2 and 3, in particular the latter. Of course, in the absence of a concrete research design, both assumptions are exceedingly strong. But these assumptions may be more plausible when inferences are based on a GQE, where comparisons are made between units on one side or the other of the boundary, perhaps after conditioning on predetermined covariates. In a succesful GQE without interference, in a sufficiently small neighborhood around the boundary, we would have $\mathbb{E}[Y_i(0)|D_i = 0, \boldsymbol{X}] = \mathbb{E}[Y_i(0), \boldsymbol{X}]$, as stated in Assumption 2. But the random or as-if random assignment of units to treated or control areas does not imply Assumption 3, the assumption on which our derivations under interference were based. We now offer some discussion on the scenarios under which the assumption can be expected to hold.

If we consider the geographic location of every unit as fixed, an experiment where every unit has the same probability of receiving treatment might result in each unit having a different probability of receiving spillovers. This arises from the fact that when the locations of units are fixed, units that are spatially isolated and have no other units near them may have a small or 0 probability of receiving spillovers. In contrast, units that are in close proximity to other units may have a positive and larger probability of receiving spillovers. In this case, estimation of the parameters of interest may require weighting the observations according to each unit's probability of receiving spillovers (Gerber and Green 2012, Ch. 8).

However, keeping units' locations fixed and randomly locating the boundary may not be the most plausible way to conceptualize treatment assignment in our application. In many geographic designs, we might view the boundary as fixed, but we assume that, within a narrow band around this boundary, units randomly choose their geographic location. Units that happen to choose a location in the treated area receive the treatment, and units that happen to choose a location within the control area receive the control condition. In this sense, the assignment of treatment is seen as a result of units' location decisions, and therefore units' locations are not seen as fixed. Under this assignment mechanism, if every unit is equally likely to select any location within a fixed band around the boundary, each unit is equally likely to receive spillovers and *ex ante*, within the fixed band, the units' potential outcomes with and without interference are equally likely to be revealed.

Importantly, if we think this form of assignment mechanism is in operation, Assumption 3 holds naturally, provided the band around the boundary is sufficiently narrow. Why might we think this is true in the context of the quasi experiment in Colorado? First, the boundary is a county border, which has existed for decades, so it is natural to think of the boundary as fixed and individual location decisions near the border as random. Secondly, our focus on a town where residents on each side of the county border share the same city amenities, gives us the basis to assume that the choice of residence on each side of the boundary is unrelated to the (control) turnout potential outcomes, once we condition on predetermined covariates. Nonetheless, we must emphasize that however plausible this form of assignment mechanism appears to be, this is an untestable and strong assumption, and our conclusions about the extent of interference depend on its validity.

## 5. Exploring Interference Effects and the Extent of Interference

Before turning to the exploration of interference in our application, we use the framework introduced above to explore some features of the interference pattern in more detail. We explore two specific issues. First, we relax the assumption that all individuals in the interference area receive interference. So far, our framework assumed that all treated individuals near enough control individuals—that is, treated individuals with $g_i(\boldsymbol{D}; \delta, \eta) = 1$—received spillovers, and all other treated individuals were free of interference. We consider a generalization of this condition where only a proportion of the treated individuals near populated control areas receive interference—while individuals with $g_i(\boldsymbol{D}; \delta, \eta) = 0$ are still assumed to receive no interference. Secondly, we consider how the results from an analysis that mistakenly ignored interference and proceeded to simply compare the outcomes among treated and control units as if there were no spillovers would differ from the interference-free effect in an analysis that took interference into account.

### 5.1 Sensitivity of interference effect when geographic spillovers affect a subset of units

Above, the effect of interference was captured by the parameter $\tau_{S,\delta\eta}$, which assumed that every individual in the interference area was in fact affected by control individuals. We now investigate what happens when only a fraction of the individuals who reside in the interference area are affected by spillovers.

We assume that treated individuals with $g_i(\boldsymbol{D}; \delta, \eta) = 1$ receive interference from the control area with probability $0 < q \leq 1$ instead of with certainty. For given values of $\delta$ and $\eta$, the expected observed outcome for treated individuals with $g_i(\boldsymbol{D}; \delta, \eta) = 1$ is now

$$\begin{aligned}\mathbb{E}[Y_i|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, \boldsymbol{X}] = \; & q \cdot \mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, \boldsymbol{X}\right] \\ & + (1-q) \cdot \mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, \boldsymbol{X}\right].\end{aligned}$$

Under this generalization, $\mathbb{E}[Y_i|D_i=1, g_i(\boldsymbol{D};\delta,\eta)=1, \boldsymbol{X}] \neq \mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i=1, g_i(\boldsymbol{D};\delta,\eta)=1, \boldsymbol{X}\right]$, and $\tau_{S,\delta\eta}(\boldsymbol{X})$ can no longer be identified. However, under Assumptions 2 and 3, we have:

$$\begin{aligned}\mathbb{E}[Y_i|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, \boldsymbol{X}] &= q\mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, X\right] \\ &\quad + (1-q)\mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, \boldsymbol{X}\right] \\ &= q\mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i = 1, \boldsymbol{X}\right] + (1-q)\mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, \boldsymbol{X}\right], \quad \text{and} \\ \mathbb{E}[Y_i|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 0|\boldsymbol{X}] &= \mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 0, \boldsymbol{X}\right] = \mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, \boldsymbol{X}\right],\end{aligned}$$

leading to

$$\begin{aligned}&\mathbb{E}[Y_i|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, \boldsymbol{X}] - \mathbb{E}[Y_i|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 0, \boldsymbol{X}] \\ &\quad = q\mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i = 1, \boldsymbol{X}\right] + (1-q)\mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, \boldsymbol{X}\right] \\ &\qquad - \mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, \boldsymbol{X}\right] \\ &\quad = q\left\{\mathbb{E}\left[Y_{i;\delta\eta}(1,1)|D_i = 1, \boldsymbol{X}\right] - \mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, \boldsymbol{X}\right]\right\} \\ &\quad = q\tau_{S,\delta\eta}(\boldsymbol{X})\end{aligned}$$

Therefore, the interference effect, $\tau_{S,\delta\eta}$, is now a function of $q$. This effect, which we denote by $\tau^q_{S,\delta\eta}(\boldsymbol{X})$, is given by:

$$\tau^q_{S,\delta\eta}(\boldsymbol{X}) = (1/q)(\mathbb{E}[Y_i|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 1, \boldsymbol{X}] - \mathbb{E}[Y_i|D_i = 1, g_i(\boldsymbol{D}; \delta, \eta) = 0, \boldsymbol{X}]).$$

Thus, whereas before all treated individuals in close proximity to (enough) control individuals were assumed to receive spillovers, now only $q$% of them have their outcomes affected by interference from individuals in the control area, while the remaining $(1 - q)$% have the same outcome they would have had if they had been in the interference-free area where $g_i(\boldsymbol{D}; \delta, \eta) = 0$.

This analysis reveals that the differences in observed average treated outcomes between the interference-free and the interference areas are equal to the interference effect $\tau^q_{S,\delta\eta}$ when $q = 1$. But when $q < 1$, the true effect of interference will be larger than the observed difference in outcomes by a factor $1/q > 1$. Thus, assuming that interference affects all units in the interference area ($q = 1$) gives a lower bound on the interference effect.

### 5.2 Characterizing the extent of interference

We now investigate the degree to which the conclusions from an analysis that ignored interference when interference was in fact present would lead to incorrect conclusions. Recall that, in the absence of interference, we defined the parameter $\tau = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]$. This parameter, however, is undefined in the presence of interference. When we allowed for interference, we focused instead on the parameters $\tau_{T,\delta\eta} = \mathbb{E}\left[Y_{i;\delta,\eta}(1,0) - Y_i(0)|D_i = 1\right]$ and $\tau_{S,\delta\eta} = \mathbb{E}\left[Y_{i;\delta,\eta}(1,1) - Y_{i;\delta,\eta}(1,0)|D_i = 1\right]$.

If units interfered with each other but an analyst made the incorrect assumption that the study is interference-free, the analyst would proceed to estimate the mean outcome differences between all units in the treatment area and all units in the control area—after conditioning on pretreatment covariates if the assumptions introduced above were invoked. Under our setup, however, the average among treated outcomes conditional on $\boldsymbol{X}$ would not equal $\mathbb{E}[Y_i(1)|D_i = 1, \boldsymbol{X}]$ but rather the weighted average of $\mathbb{E}\left[Y_{i;\delta,\eta}(1,1)|D_i = 1, \boldsymbol{X}\right]$ and $\mathbb{E}\left[Y_{i;\delta,\eta}(1,0)|D_i = 1, \boldsymbol{X}\right]$. Letting $p_{\mathrm{I}} = \Pr[g_i(\boldsymbol{D}; \delta, \eta) = 1]$, the estimand estimated by the analyst that ignored interference would be

$$\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1, \boldsymbol{X}] - \mathbb{E}[Y_i|D_i = 0, \boldsymbol{X}] = \\
&\qquad = \{p_{\mathrm{I}} \cdot \mathbb{E}[Y_i|D_i = 1, \boldsymbol{X}, g_i(\boldsymbol{D}; \delta, \eta) = 1] + (1 - p_{\mathrm{I}}) \cdot \mathbb{E}[Y_i|D_i = 1, \boldsymbol{X}, g_i(\boldsymbol{D}; \delta, \eta) = 0]\}\,. \\
&\qquad\quad - \mathbb{E}[Y_i(0)|D_i = 1, \boldsymbol{X}]
\end{aligned}$$

The estimand is now a comparison between the control units and a weighted average of treated units—some of which are subject to interference and some not. Writing the overall estimand this way we observe that, when $p_{\mathrm{I}}$ is small, the approach that ignored interference and pooled all treated observations would closely approximate the interference-free treatment effect, $\tau_{T,\delta\eta}$, since in this case, under the assumptions of our framework, we would have:

$$\begin{aligned}
&p_{\mathrm{I}} \cdot \mathbb{E}[Y_i|D_i = 1, \boldsymbol{X}, g_i(\boldsymbol{D}; \delta, \eta) = 1] + (1 - p_{\mathrm{I}}) \cdot \mathbb{E}[Y_i|D_i = 1, \boldsymbol{X}, g_i(\boldsymbol{D}; \delta, \eta) = 0] \\
&\approx \mathbb{E}[Y_i|D_i = 1, \boldsymbol{X}, g_i(\boldsymbol{D}; \delta, \eta) = 0] = \mathbb{E}\left[Y_{i;\delta\eta}(1,0)|D_i = 1, \boldsymbol{X}\right],
\end{aligned},$$

the first term in $\tau_{T,\delta\eta}$.

Moreover, when the potential outcomes are bounded as they are in our voting application where they take values equal to 0 (non voting) or 1 (voting), given a value of $p_{\mathrm{I}}$, we can calculate the maximum value of the term $p_{\mathrm{I}} \cdot \mathbb{E}[Y_i | D_i = 1, \boldsymbol{X}, g_i(\boldsymbol{D}; \delta, \eta) = 1]$. In our case, this upper bound is $p_{\mathrm{I}}$.

This discussion illustrates that it is not only of interest to calculate the effect of interference but also to establish whether interference affects a large enough proportion of units. If it does not—that is, if $p_{\mathrm{I}}$ is small—in the case of bounded outcomes, we may be able to assert that an analysis that ignored interference when in fact interference is present would produce treatment effect estimates that would approximate the effects under no interference.

## 6. Application to the 2010 Colorado Primary

We now apply the framework described above to analyze the 2010 Colorado primary, starting with a description of the data. Our main source of information is the Colorado voter registration file, the database of registered voters maintained by the state of Colorado for administrative purposes. We acquired these data from a private vendor. The administrative data from the state contain a limited number of covariates including date of birth, gender, voting history, voters' addresses, and the legislative districts in which each voter's address is included. The private vendor also includes an additional variable estimating the voter's likely race. In this region, most voters are White with a substantial minority of Hispanic voters. To determine voter locations, we geocoded each voter's location using the address in the voter file.

We first restrict our data to include only those individuals who in 2010 lived in the central area of Basalt that is split by the border between Eagle and Pitkin counties—the border that determines all-mail or in-person voting. Within the central area of Basalt, our covariate-adjusted results condition on the set of pre-treatment covariates that we have available: age, gender, whether the individual is Hispanic, voting history for 2008 and 2006, party affiliation as declared in the registration file, and an indicator for whether the individual's registration status is considered active by the state. Since our main data source is the Colorado registration file, all our turnout measures—including both pre-treatment turnout shares and the turnout share in the 2010 primary—are constructed as the proportion of individuals in the registration file that turn out to vote in a given election. This means that our turnout measures condition on registration status. We discuss the potential methodological complications associated with such measures in Section 6.2.

Before presenting the estimation results, we examine the observed covariates in our sample within central Basalt, to asses whether the treated and control areas inside this small region are already comparable in terms of these characteristics. Table 1 contains sample means and the absolute standardized differences in means (difference in means divided by the pooled standard deviation between groups before matching) for three demographic characteristics, voter registration status, and turnout in the last four elections. While geographic proximity produces acceptable balance on residents' Hispanic ethnicity, gender, age, and active registration status, there are larger differences in turnout in past elections, with standardized differences in these variables exceeding 0.20. These differences suggest that unadjusted comparisons between the groups cannot be interpreted as causal effects of all-mail elections on 2010 voter turnout.

**Table 1.** Covariate balance between treated and control areas around the geographic border

| | Mean treated | Mean control | Std. diff. |
|---|---|---|---|
| Hispanic | 0.07 | 0.08 | 0.03 |
| Age | 48.4 | 45.6 | 0.19 |
| Female | 0.49 | 0.50 | 0.01 |
| Active Registration | 0.24 | 0.29 | 0.10 |
| 2008 General Election Turnout | 0.71 | 0.60 | 0.24 |
| 2008 Primary Election Turnout | 0.07 | 0.02 | 0.27 |
| 2006 General Election Turnout | 0.48 | 0.36 | 0.24 |
| 2006 Primary Election Turnout | 0.06 | 0.01 | 0.22 |

*Notes*: Total sample size is 977 treated (all-mail) voters and 620 control (in-person) voters. Std. diff. = absolute standardized difference. Means for turnout are proportion of registered individuals voting in that election.

**Table 2.** Estimated average effect of all-mail voting on turnout on treated area under no interference, 2010 Colorado primary election

| | Unadjusted | Covariate-adjusted |
|---|---|---|
| Difference in turnout rates | 0.010 | −0.059 |
| 95% confidence interval | (−0.027, 0.047) | (−0.091,−0.027) |
| Control turnout rate | 0.155 | |

*Notes*: Total sample size is 977 treated (all-mail) voters and 620 control (in-person) voters. Turnout shares calculated as proportion of registered individuals voting in that election.

We estimate Equation (2) by least-squares using the variables in Table 1 as covariates. Table 2 contains point estimates of $\tau$ and associated 95% confidence intervals. We also report the unadjusted least-squares estimator of $\tau$ corresponding to Equation (1), which is simply the unadjusted difference in turnout rates between treated and control areas. This unadjusted estimate is reported only for completeness, but we do not believe it can be interpreted as a causal effect due to the observable pre-treatment differences between the areas.

We find that, in a narrow band around the boundary between the treated and control areas and after conditioning on a set of observed characteristics, the voter turnout rate was 6 percentage points lower in the all-mail county, with a 95% confidence interval ranging from −0.091 to −0.027. Given our assumptions, this difference is the average all-mail voting effect on turnout for the residents in the treated area. This analysis assumes that outcomes of an individual do not depend on the treatment status of other individuals, a constraint that we relax in the following section.

## 6.1 Empirical results allowing for spatial interference

We now re-analyze the effects above based on our GQE. Under our setup and given the assumptions introduced above, we calculate the interference-free treatment, $\tau_{\mathrm{T},\delta\eta}$, and the interference effect, $\tau_{\mathrm{T},\delta\eta}$, for different values of $\eta$ and $\delta$.

To implement estimation of these quantities, we calculate, given $\delta$ and $\eta$, the interference set, $\mathcal{I}_{\delta\eta}$, which is simply the collection of all treated individuals in our data for whom

$g_i(\boldsymbol{D}; \delta, \eta) = 1$. Once we form $\mathcal{I}_{\delta\eta}$, we can estimate $\tau_{S,\delta\eta}$ through a comparison of treated individuals that are in $\mathcal{I}_{\delta\eta}$ to those who are not, since this set identifies all treated individuals that are spatially proximate to enough control voters and may have been subject to interference. Similarly, we estimate $\tau_{T,\delta\eta}$ through a comparison of control voters to treated voters not in $\mathcal{I}_{\delta\eta}$. To form $\mathcal{I}_{\delta\eta}$, we must locate the distance between each treated unit and each control unit, calculate whether any control units reside within $\delta$ meters of each treated unit, and then count the number of control units within $\delta$ distance. We calculate this set using the following algorithm.

**Algorithm 6.1:** Computing the Interference Set($\mathcal{I}_{\delta\eta}$,$\delta$,$\eta$)

**for** $i \leftarrow 1$ to $m$ treated units

**do** $\begin{cases} \text{Calculate distance from treated } i \text{ to all controls} \\ \text{Locate all controls within } \delta \text{ distance of treated } i \\ N_{i\delta} \leftarrow \text{ number of control units that are } \delta \text{ distance from treated } i \\ \text{Place treated unit } i \text{ in } \mathcal{I}_{\delta\eta} \text{ if } N_{i\delta} \geq \eta \end{cases}$

**return** $(\mathcal{I}_{\delta\eta})$

The size of $\mathcal{I}_{\delta\eta}$ depends on the values we choose for $\delta$ and $\eta$. As we make $\delta$ larger and $\eta$ smaller, we allow for interference to become more severe. Once we have formed the set $\mathcal{I}_{\delta\eta}$, calculation of the quantities of interest is straightforward.

We apply these methods to the data from the 2010 primary in Colorado. For a set of $\eta$ and $\delta$ values, we estimate $\tau_{S,\delta\eta}$ and $\tau_{T,\delta\eta}$. Different values of $\eta$ and $\delta$ allow for more or less stringent definitions of spatial interference. For example, if we set $\delta$ to 250 m and $\eta$ to 1, our model of interference asserts that all treated units who have at least one control unit within a 250-m radius of their location were subject to interference. We set $\delta$ to 250 and 100 m, and $\eta$ to 1, 5, and 10 individuals. We estimate $\tau_{S,\delta\eta}$ and $\tau_{T,\delta\eta}$ for each set of values to observe whether these quantities change as a function of both distance and the density of control units that treated individuals are near to.

Table 3 contains estimates for $\tau_{S,\delta\eta}$ and $\tau_{T,\delta\eta}$ (which we denote as $\hat{\tau}_{S,\delta\eta}$ and $\hat{\tau}_{T,\delta\eta}$) along with 95% confidence intervals for each set of $\eta$ and $\delta$ values. The table also reports the proportion of treated voters in the interference set (i.e., an estimate of $p_I$). All results in the table are covariate-adjusted using least-squares estimates from Equation (3), employing heteroscedasiticty-robust standard errors.

First, we consider the estimates when $\delta = 100$ (top panel). When $\eta = 1$, we let interference affect all treated units who live within 100 m of at least one control unit. Under this interference scenario, 9.6% of the treated observations are affected by interference, and estimate of $\tau_{T,\delta\eta}$ is $-0.059$ (95% confidence interval from $-0.092$ to $-0.026$), which is very similar to our covariate-adjusted estimate of $\tau$ in Table 2—the ATT estimated under no interference. In contrast, the estimate for $\tau_{S,\delta\eta}$ is $-0.0056$, much closer to 0 and not significantly different from 0 (95% confidence interval ranges from $-0.061$ to 0.050). Thus, for $\delta = 100$ and $\eta = 1$, there is not a statistically significant difference in the patterns of voter turnout between the interference and the interference-free area. When $\eta$ is 5 or 10, the proportion of treated voters inside the interference area is naturally smaller. In this case, the estimates of $\tau_{S,\delta\eta}$ are, respectively, $-0.012$ and $-0.054$, both indistinguishable from 0. Since

**Table 3.** Estimated average effects of all-mail voting on turnout on treated area under difference interference scenarios, 2010 Colorado primary election

| | $\eta = 1$ | $\eta = 5$ | $\eta = 10$ |
|---|---|---|---|
| | | $\delta = 100$ meters | |
| Interference-free effect ($\hat{\tau}_{T,\delta\eta}$) | −0.059 | −0.059 | −0.058 |
| | (−0.092, −0.026) | (−0.091, −0.027) | (−0.090, −0.026) |
| Interference effect ($\hat{\tau}_{S,\delta\eta}$) | −0.0056 | −0.012 | −0.054 |
| | (−0.061, 0.050) | (−0.109, 0.085) | (−0.155, 0.047) |
| $\hat{p}_{\mathrm{I}}$ | 0.096 | 0.045 | 0.031 |
| | | $\delta = 250$ meters | |
| Interference-free effect ($\hat{\tau}_{T,\delta\eta}$) | −0.069 | −0.062 | −0.058 |
| | (−0.104, −0.034) | (−0.095, −0.029) | (−0.091, −0.025) |
| Interference effect ($\hat{\tau}_{S,\delta\eta}$) | 0.033 | 0.012 | −0.011 |
| | (−0.0078, 0.0738) | (−0.037, 0.061) | (−0.060, 0.038) |
| $\hat{p}_{\mathrm{I}}$ | 0.294 | 0.177 | 0.156 |

*Notes*: Total sample size is 977 treated (all-mail) voters and 620 control (in-person) voters. Turnout shares calculated as proportion of registered individuals voting in that election.

the proportion of treated units in the interference set is very small ($p_{\mathcal{I}}$ is equal to 0.045 and 0.031 for $\eta = 5$ and $\eta = 10$, respectively), the covariate-adjusted estimate for $\tau_{\mathrm{T},\delta\eta}$ remains very similar to the analogous estimate for $\eta = 1$—and also to the covariate-adjusted effect under no interference reported in Table 2.

Next, we consider the case of $\delta = 250$ (bottom panel of Table 3). When $\eta = 1$, the estimated interference-free effect is −0.069 (confidence interval ranging from −0.104 to −0.034), similar to our previous estimates. In addition, turnout is about 3.3 percentage points higher for treated voters in the interference area relative to the turnout of treated voters in the interference-free area, but this effect is again statistically indistinguishable from 0. Note that in this scenario interference is assumed to be much more prevalent, with 29% of treated voters inside the interference area. When we set $\eta$ to either 5 or 10, the estimated effect of interference, $\hat{\tau}_{\mathrm{S},\delta\eta}$, continues to be indistinguishable from 0. In contrast, the confidence intervals including for $\tau_{\mathrm{T},\delta\eta}$ are $[-0.095, -0.029]$ for $\eta = 5$ and $[-0.091, -0.025]$ for $\eta = 10$, very similar to the confidence intervals under $\delta = 100$ and also to confidence intervals for $\tau$ ignoring interference reported in Table 2.

In sum, our analysis shows that, under our assumptions, the interference effect is indistinguishable from 0 in all cases. Moreover, except under the most extreme scenario of interference with $\delta = 1$ and $\eta = 250$ where 29% of treated voters are in the interference set, the proportion of treated voters in the interference set tends to be small (between 17 and 3%), leading to an interference-free effect that is similar to the effect estimated assuming that interference is not present. These results suggest that all-mail voting reduced voter turnout in the 2010 primary election by an average of about 6 percentage points.

## 6.2 Sensitivity to differential voter registration

A potential complication with our analysis is that our data source is the Colorado voter registration file, and thus our measure of voter turnout is constructed as the proportion of registered citizens who turn out to vote. If the decision to register is itself affected by the mode of voting, our reported results could misrepresent the true turnout effects. Eagle county's decision to adopt all-mail voting was announced in 2010 before the primary election was held, making it possible for citizens to adjust their election registration decisions in response to the upcoming change in the mode of voting.

The ideal solution would be to obtain a list of the total voting eligible population in the treated and control areas at the moment of the 2010 primary election. Unfortunately, such data are unavailable. An alternative is to follow the approach in Nyhan et al. (2017) and explore how much differential registration between the treated and control groups could occur before our observed turnout effects—which construct turnout shares as total voters over total registration—became consistent with a zero effect on the true turnout share—the share of voters to the total voting eligible population. Generalizing the sensitivity analysis in Nyhan et al. (2017) to include covariates, for every estimated effect reported in Table 3, we report the differential registration factor $k^{\star}$—the amount of differential registration between treated and control groups that would be required to produce the difference in turnout-to-registration rates we observed if the true turnout effects were equal to 0.

Given a treated and a control or reference group, the differential registration factor is estimated by simply dividing the turnout-to-registration share in the control group ($T_{\mathrm{c}}^{\mathrm{Reg}}$) over the turnout-to-registration share in the treatment group ($T_{\mathrm{t}}^{\mathrm{Reg}}$). For estimation of the $k^{\star}$ associated with $\tau_{T,\delta\eta}$, $T_{\mathrm{t}}^{\mathrm{Reg}}$ includes all units in the treatment group outside of the interference region ($D_i = 1$ and $g_i(\boldsymbol{D}; \delta, \eta) = 0$), and $T_{\mathrm{c}}^{\mathrm{Reg}}$ includes all control units ($D_i = 0$). For estimation of the $k^{\star}$ associated with $\tau_{S,\delta\eta}$, $T_{\mathrm{t}}^{\mathrm{Reg}}$ includes all units in the treatment group inside the interference region ($D_i = 1$ and $g_i(\boldsymbol{D}; \delta, \eta) = 1$), and $T_{\mathrm{c}}^{\mathrm{Reg}}$ includes units in the treatment group outside of the interference region ($D_i = 0$ and $g_i(\boldsymbol{D}; \delta, \eta) = 0$). To incorporate covariates, we estimate $T_{\mathrm{c}}^{\mathrm{Reg}}$ with the average predicted values from the linear model for all observations in the treatment group, but with the corresponding treatment indicator set to 0.

The results of the sensitivity analysis are shown in Table 4, where we report, for each of the three values of $\eta$ combined with $\delta = 100$ or $\delta = 250$, the differential registration factor for both $\tau_{T,\delta\eta}$ and $\tau_{S,\delta\eta}$. For each combination of $\eta$ and $\delta$, the first two columns report $T_{\mathrm{c}}^{\mathrm{Reg}}$ and $T_{\mathrm{t}}^{\mathrm{Reg}}$, the estimated values of the turnout-to-registration shares for each group—the difference between these values are the effects reported in Table 3. The third column reports the differential registration factor associated with each effect. For example, for $\eta = 1$ and $\delta = 250$, $k^{\star}$ is 1.447, showing that the rate of registration in the treatment group would have

**Table 4.** Sensitivity of average effects of all-mail voting to differential registration, 2010 Colorado primary election

| | $\eta = 1$ | | | $\eta = 5$ | | | $\eta = 10$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $T_{\mathrm{t}}^{\mathrm{Reg}}$ | $T_{\mathrm{c}}^{\mathrm{Reg}}$ | $k^{\star}$ | $T_{\mathrm{t}}^{\mathrm{Reg}}$ | $T_{\mathrm{c}}^{\mathrm{Reg}}$ | $k^{\star}$ | $T_{\mathrm{t}}^{\mathrm{Reg}}$ | $T_{\mathrm{c}}^{\mathrm{Reg}}$ | $k^{\star}$ |
| | | | | $\delta = 100$ meters | | | | | |
| Interference-free effect ($\hat{\tau}_{T,\delta\eta}$) | 0.165 | 0.224 | 1.355 | 0.165 | 0.224 | 1.355 | 0.166 | 0.224 | 1.346 |
| Interference effect ($\tau_{S,\delta\eta}$) | 0.117 | 0.123 | 1.048 | 0.182 | 0.194 | 1.066 | 0.133 | 0.188 | 1.407 |
| | | | | $\delta = 250$ meters | | | | | |
| Interference-free effect ($\hat{\tau}_{T,\delta\eta}$) | 0.155 | 0.224 | 1.447 | 0.163 | 0.224 | 1.378 | 0.166 | 0.224 | 1.345 |
| Interference effect ($\tau_{S,\delta\eta}$) | 0.171 | 0.137 | 0.805 | 0.156 | 0.144 | 0.921 | 0.132 | 0.143 | 1.083 |

*Notes*: Total sample size is 977 treated (all-mail) voters and 620 control (in-person) voters. Turnout shares calculated as proportion of registered individuals voting in that election.

to be 44.7 percentage points higher in the treatment group than in the control group to make the estimated interference free treatment effect $\hat{\tau}_{T,\delta\eta} = -0.069$ (reported in Table 3 and also obtained from Table 4 as 0.155–0.224) consistent with a 0 effect on true turnout rates. A 44.7 percentage point difference in registration rates is a very large effect, unlikely to occur in practice. This large value of $k^{\star}$ suggests that the interference-free effect is robust: the negative turnout effect would remain even with substantial differential registration between the treated and control groups.

In general, turnout-to-registration effects are more sensitive to differential registration whenever these effects are larger in absolute value and whenever the turnout-to-registration share in the baseline group is smaller. Since our estimated interference-free effects $\tau_{T,\delta\eta}$ are much larger in absolute value than the interference effects $\tau_{S,\delta\eta}$, and these effects are large relative to $T_{\mathrm{c}}^{\mathrm{Reg}}$, the pattern in Table 4 is consistent: the differential registration factor $k^{\star}$ is large for the interference-free treatment effect $\tau_{T,\delta\eta}$ in all cases, and small for the interference effects $\tau_{S,\delta\eta}$. Our conclusion is that the estimated interference effects $\tau_{S,\delta\eta}$, even if statistically distinguishable from 0, would be less robust to differential registration patterns. For example, for $\eta = 5$ and $\delta = 100$, the registration factor of 1.066 indicates that a difference in registration rates of 6.6 percentage points would be sufficient to make the observed interferecne effect $\tau_{S,\delta\eta} = -0.012$ consistent with a 0 effect on true turnout rates.

Finally, we note that all our conclusions about sensitivity to differential registration assume that the interference set is correctly calculated based on the registration file. This implies the assumption that individuals who reside in the control area and are not registered to vote do not affect the potential outcomes of treated individuals in the all-mail voting area. This may be plausible if we assume that residents who are not registered are not known to get-out-the-vote campaigns and are not actively involved in political activities.

## 7. Discussion

Policymakers seldom use randomized experiments to study the effects of different voting regulations such as modes of voter registration or convenience voting policies on voter participation. GQEs such as the one we examine here may therefore provide fruitful opportunities for researchers to study policy effects that would otherwise go unexplored. Such natural experiments are not only useful in social science applications—where the assumption of comparability on either side of the geographic border is always a strong assumption—but also in other disciplines where the potential confounders are more closely related to the physical characteristics of the terrain and thus more likely to be offset by spatial proximity (see Wonkka et al. 2015).

In our analysis of the 2010 Colorado primary, we find that vote-by-mail elections appear to suppress turnout. Using our framework, we find that, unless we assume a fairly strong patten of interference, the treatment effects estimated under a framework that ignores interference would not differ from the interference-free effect in our spatial interference framework. Moreover, our estimated interference effect could not be distinguished from 0 in any of the scenarios we considered. Thus, given our assumptions, interference between voters does not seem to be prevalent in our application, as was also found by Sinclair et al. (2012).

The prior literature has found both positive and negative effects of all-mail reforms on voter turnout; our results are consistent with prior evidence of negative effects. The reasons behind the different conclusions across studies are hard to ascertain, as all studies, including our own, are non-experimental and potentially threatened by unobserved confounders driving the decision to adopt voting reforms. In addition to these threats to internal validity, the differences could be partially due to heterogeneity in the trade-off between convenience voting and the social aspects of voting in different political contexts. In those places where voting barriers affect a large proportion of potential voters, the lower barriers to voting as a result of convenience voting reforms may more than compensate the lower turnout that may be induced by removing the social aspect of voting. In contrast, in settings where most citizens can afford the costs of voting in the absence of convenience reforms, removing the social aspect and focal point of election day may lead to a decrease in turnout that is not compensated by making voting easier. Since we focus on a primary election, we are focusing on a subset of citizens who are highly committed and vote in most elections; moreover, in the 2010 Colorado primary, voters had three salient races on the ballot. This suggests that the scenario we study might belong to the latter category, where the elimination of the social and shared aspects of voting is relatively more costly.

## Acknowledgements

This manuscript was prepared for the CESifo Economic Studies special issue on geocoding to follow our keynote lecture presentation at the 2016 CESifo Economic Studies Conference on the Use of Geo-coded Data in Economic Research, held in Munich in 18 and 19 November 2016. The authors thank Matias Cattaneo, Jake Bowers, Jeff Lewis, seminar participants at the University of California, Berkeley, and conference participants at the 2013 Causality in Political Networks Conference at the University of Chicago, the 2014 Southeastern Methods Meeting at the University of South Carolina, the 2015 Harvard University Conference on Political Geography, and the 2016 CESifo Economic Studies Conference on the Use of Geo-coded Data in Economic Research, for valuable comments and discussion. R.T. gratefully acknowledges financial support from the National Science Foundation (SES 1357561). A previous version of this manuscript was circulated under the title 'A Framework for Understanding Interference Between Units In Geographic Natural Experiments with an Application to All-Mail Voting in Colorado'.

## References

Aronow, P. M. and C. Samii (2017), "Estimating Average Causal Effects Under Interference Between Units," *Annals of Applied Statistics* **11**, 1921–047.

Bergman, E. and P. A. Yates (2011), "Changing Election Methods: How Does Mandated Vote-by-Mail Affect Individual Registrants?", *Election Law Journal* **10**, 115–27.

Berinsky, A. J., N. Burns, and M. W. Traugott (2001), "Who Votes by Mail?: A Dynamic Model of the Individual-Level Consequences of Voting-by-Mail Systems", *Public Opinion Quarterly* **65**, 178–97.

Bowers, J., M. M. Fredrickson, and C. Panagopoulos (2013), "Reasoning about Interference between Units: A General Framework", *Political Analysis* **21**, 97–124.

Galiani, S., P. J. McEwan, and B. Quistorff (2017), *Regression Discontinuity Designs: Theory and Applications, Advances in Econometrics*, vol. **38**. In M. D., Cattaneo and J. C., Escanciano eds, Emerald Publishing Limited, Bingley, United Kingdom, pp. 195–236.

Gerber, A. S. and D. P. Green (2012), *Field Experiments: Design, Analysis, and Interpretation*. Norton, New York, NY.

Gerber, A. S., D. P. Green, and C. W. Larimer (2008), "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment", *American Political Science Review* **102**, 33–48.

Gerber, A. S., G. A. Huber, and S. J. Hill (2013), "Identifying the Effect of All-Mail Elections on Turnout: Staggered Reform in the Evergreen State", *Political Science Research and Methods* **1**, 91–116.

Gronke, P., E. Galanes-Rosenbaum, P. A. Miller, and D. Toffey (2008), "Convenience Voting", *Annual Review of Political Science* **11**, 437–55.

Gronke, P. and P. Miller (2012), "Voting by Mail and Turnout in Oregon Revisiting Southwell and Burchett", *American Politics Research* **40**, 976–97.

Holland, P. W. (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association* **81**, 945–60.

Hong, G. and S. W. Raudenbush (2006), "Evaluating Kindergarten Retention Policy", A Case Study of Causal Inference for Multilevel Observational Data," *Journal of the American Statistical Association* **101**, 901–10.

Hudgens, M. G. and M. E. Halloran (2008), "Toward Causal Inference with Interference", *Journal of the American Statistical Association* **103**, 832.

Karp, J. A. and S. A. Banducci (2000), "Going Postal: How All-Mail Elections Influence Turnout", *Political Behavior* **22**, 223–39.

Keele, L., S. Lorch, M. Passarella, D. Small, and R. Titiunik (2017), "An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance," in M. D., Cattaneo and J. C., Escanciano, eds, *Regression Discontinuity Designs: Theory and Applications, Advances in Econometrics*, vol. **38**. Emerald Publishing Limited, Bingley, United Kingdom, pp. 147–94.

Keele, L. and R. Titiunik (2015), "Geographic Boundaries as Regression Discontinuities", *Political Analysis* **23**, 127–55.

Keele, L., R. Titiunik, and J. Zubizarreta (2015), "Enhancing a Geographic Regression Discontinuity Design through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout", *Journal of the Royal Statistical Society: Series A* **178**, 223–39.

Keele, L. J. and R. Titiunik (2016), "Natural Experiments Based on Geography", *Political Science Research and Methods* **4**, 65–95.

Kousser, T. and M. Mullin (2007), "Does Voting by Mail Increase Participation? Using Matching to Analyze a Natural Experiment", *Political Analysis* **15**, 428–45.

Nyhan, B., C. Skovron, and R. Titiunik (2017), "Differential Registration Bias in Voter File Data: A Sensitivity Analysis Approach", *American Journal of Political Science* **61**, 744–60.

Richey, S. (2008), "Voting by Mail: Turnout and Institutional Reform in Oregon", *Social Science Quarterly* **89**, 902–15.

Rosenbaum, P. R. (2007), "Interference between Units in Randomized Experiments", *Journal of the American Statistical Association* **102**, 191–200.

Rubin, D. B. (2005), "Causal Inference Using Potential Outcomes", *Journal of the American Statistical Association* **100**, 322.

Sinclair, B., M. McConnell, and D. P. Green (2012), "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments", *American Journal of Political Science* **56**, 1055–69.

Sobel, M. E. (2006), "What Do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in The Face of Interference," *Journal of the American Statistical Association* **101**, 1398–407.

Southwell, P. L. and J. I. Burchett (2000), "The Effect of All-Mail Elections on Voter Turnout", *American Politics Research* **28**, 72–9.

Tchetgen Tchetgen, E. J. and T. J. VanderWeele (2012), "On Causal Inference in the Presence of Interference", *Statistical Methods in Medical Research* **21**, 55–75.

Vanderweele, T. (2008), "Ignorability and Stability Assumptions in Neighborhood Effects Research", *Statistics in Medicine* **27**, 1934–43.

VanderWeele, T. J., G. Hong, S. M. Jones, and J. L. Brown (2013), "Mediation and Spillover Effects in Group Randomized Trials: A Case Study of the 4Rs Educational Intervention", *Journal of the American Statistical Association* **108**, 469–81.

Verbitsky-Savitz, N. and S. W. Raudenbush (2012), "Causal Inference under Interference in Spatial Settings: A Case Study Evaluating Community Policing Program in Chicago", *Epidemiologic Methods* **1**.

Verdery, A. M., B. Entwisle, K. Faust, and R. R. Rindfuss (2012), "Social and Spatial Networks: Kinship Distance and Dwelling Unit Proximity in Rural Thailand", *Social Networks* **34**, 112–27.

Wonkka, C. L., W. E. Rogers, and U. P. Kreuter (2015), "Legal Barriers to Effective Ecosystem Management: Exploring Linkages between Liability, Regulations, and Prescribed Fire", *Ecological Applications* **25**, 2382–93.

*CESifo Economic Studies*, 2018, 255–291
doi: 10.1093/cesifo/ify002
Advance Access Publication Date: 5 March 2018
Original article

# Take Me to the Centre of Your Town! Using Micro-geographical Data to Identify Town Centres

**Paul Cheshire*, Christian Hilber†, Piero Montebruno‡ and Rosa Sanchis-Guarner‖**

*Department of Geography and Spatial Economics Research Centre (SERC/CEP), London School of Economics. e-mail: p.cheshire@lse.ac.uk, †Department of Geography and Spatial Economics Research Centre (SERC/CEP), London School of Economics. e-mail: c.hilber@lse.ac.uk, ‡Department of Geography, University of Cambridge. e-mail: pfm27@cam.ac.uk and ‖Imperial College Business School and SERC/CEP, London School of Economics. e-mail: r.sanchis-guarner@imperial.ac.uk

## Abstract

We often talk about 'Town Centres' (TCs), but defining their location and extent is surprisingly difficult. Their boundaries are hard to pin down and intrinsically fuzzy. Nevertheless, policymakers often speak or act as if their definition was self-evident. The Dutch and later the British governments, for example, introduced very specific policies for them without ever clearly defining what or where they were. In this article, we propose a simple methodology to predict TC boundaries and extent. Using a range of micro-geographical data, we test our method for the whole of Great Britain in an attempt to capture all the dimensions of 'town centredness' in a 3D surface. We believe this is a contribution in its own right but is also an essential step if there is to be any rigorous analysis of TC or evaluation of policies directed at them. Our method should contribute to improve not just debates about cities, shopping hierarchies, and TCs but also to other more general debates where people and policy proceed ahead of any clear definition of what are the objects of interest. (JEL codes: L81, R12, R52)

**Key words**: town centre, planning, retail sector, land use

## 1. Introduction

Imagine you are anywhere in a city—London, Lyon, Berlin, and Wolverhampton—and you know that city well. Suddenly, someone comes up to you and asks, 'Could you tell me where the town centre is?' This could appear to be a simple, even a trivial, question, but it is not. In fact, in many instances, it proves to be surprisingly hard to answer. The aim of this article is to devise a method which could provide a response and not just a response but an answer which meets the criterion of being replicable. If you apply the method to a different town, your answer will be strictly comparable, and you would get the same answer asking different people so long as they applied the method.

This question has a particular salience, since, in many countries, there are influential urban policies that apply to 'Town Centres' (TCs).[1] But, if we cannot define the boundaries of these areas, not only can we not identify the actual areas the policies are supposed to apply to, we cannot evaluate any effects such 'Town Centre policies' may have on outcomes. Our aim in this article is to design, explain, apply, and test a method to answer this apparently trivial question. We are not concerned with *why* the TC is sought. Instead, we explore and provide an operational answer. We do this in the specific context of Britain but would suggest both the question and our approach have significant application elsewhere.

Our interest in identifying and predicting TC space arose as one part of an investigation into the effects of 'Town Centre First Policy' (TCFP) on shoppers' travel patterns (Cheshire et al. 2017), as adopted in England in 1996 (Department of the Environment 1996). This policy, remarkably similar to that applied in The Netherlands some 15 years earlier (Evers 2002), was intended to 'redirect development, not just in retailing but in all "key Town Centre uses," including leisure, office development and other uses, such as restaurants, to Town Centres', although the policy most notably affected the location of new retail development. So TC protection strengthened in England just as in the Netherlands it was becoming more flexible to support the competitiveness of the retail sector (Evers 2002). As was shown in Cheshire et al. (2015), TCFP policy did, indeed, have a substantial negative impact on total factor productivity in the English supermarket sector.

The avowed purpose of policies to support TCs was to maintain their 'viability' or, in the case of The Netherlands, to ensure that the distribution of retail outlets corresponded to the urban hierarchy. But in England TCFP was specifically introduced to facilitate 'linked shopping trips' and allow shopping trips to be undertaken using public transport—partly with the aim of reducing their carbon footprint but also for equity purposes: to protect access to shops of those without cars. Evaluation of such policies, therefore, necessarily requires information on patterns of shopping trips and changes in the extent to which shopping destinations are located in TCs.

To begin to assess the impacts of the TCFP—or any policy aimed at TCs—it is thus necessary to have definitions of where and what TCs are[2] and to be able to apply the same definitions to contexts where TCFPs were not introduced. TCFP was implemented, however, with no such definitions. While for England and Wales TCs were subsequently defined in research commissioned by the relevant government department (ODPM 2004),[3] these were not official nor are they enforced: 'It should be noted that these areas [Areas of Town Centre Activity] have no policy status and are not town centres for policy purposes – such centres will be designated in development plans' (ODPM and CASA 2002). To provide the tools for such an evaluation, the focus of the present article is to develop a method for predicting and estimating the location and extent of TC space in both England and Wales and in Scotland. In addition, we would expect our method to be widely applicable.

1 For example, Denmark, Germany, The Netherlands, or Sweden: see Guimarães (2016) for a recent survey of some of these.

2 The Scottish government produced a glossy 138 page handbook called the Town Centre Toolkit in 2015 without ever defining what a 'Town Centre' was.

3 ODPM, created in 2002 and predecessor of the DCLG.

To do this, we first obtained data on TCs as defined for 2000 from the Department for Communities and Local Government (DCLG).[4] Even with the caveat that they have ‘no policy status’, these ‘official’[5] TC definitions are the most reliable and accurate definitions of TC space in England and Wales. They consist of GIS shapefiles for 1075 TCs, of which the majority are defined as ‘Areas of Town Centre Activity’ (ATCA) and 46 as ‘Retail Cores’ (RCs)—which overlap and are sub-centres of the ATCAs. From these shapefiles we obtain the centroids of the England and Wales TCs (called DCLG TCs in what follows). This identifies the central point in each town or city. Separately, we obtain a list of alternative TCs for all Britain, from the towns and cities list in the Ordnance Survey (OS) Gazetteer and locate their central points. Below we refer to these as OSC TCs (Ordnance Survey Cities Town Centres).

To predict the extent of the TCs around these two sets of locations, we use abundant small-scale geographical information, in a range of 1–3 km from the centroids. We calculated a long list of geographical and socio-economic factors that relate to TC activities, following closely the variables used by DCLG in the construction of their Index of Town Centre Activity (ODPM 2004). To assess the extent to which these factors accurately predict TC space, we regress them on the radius of the DCLG TCs (derived from the area of the shapefiles), to replicate as closely as possible the areas of these TCs for England and Wales. We then subject the results to robustness checks and, having satisfied ourselves as to the results, apply the estimated coefficients in a separate exercise to the set of locations (OSC TCs) available for all three countries of Great Britain to predict the size of their TCs. By doing this, we obtain a full set of estimated TC boundaries for all countries in Britain, and, in particular, Scotland, on a measure consistent with that used to identify the DCLG TCs for just England and Wales.

We believe this article makes three contributions. First, we show how important it is to have clear and replicable measures of TCs to be able to consistently evaluate policies aimed at these particular locations. This is an issue which both the interested academic and policy communities seem to have surprisingly overlooked. Second, we propose a simpler methodology than others available in the literature to predict the extent of TC space around a set of locations (as discussed in Section 2). Our method requires less data than others and uses straightforward regression techniques. Finally, we provide the necessary tools to implement a robust evaluation of policies applying to TC locations, in particular for the British context where these policies are very popular with planners and policymakers.

The rest of the article is organized as follows. In Section 2 we review the existing theories relating to TCs and how those, in turn, relate to work on the urban system. Then in Section 3, we discuss the definition of TCs and some existing methods to identify their location and boundaries. In Section 4, we describe the existing data on TCs for England and Wales. In Section 5, we explain our methodology to predict the location and extent of TCs for all of Great Britain. Section 6 presents the results and provides some statistics to check how well the method works. Finally, Section 7 concludes.

4 Data for 2004 can be accessed at https://data.gov.uk/dataset/english-town-centres-2004, but we have also had access to data for years 2000 and 2002 provided to us by DCLG. These data were originally created by the ODPM. Their methodology is described in ODPM (2004).

5 As we have said, there are only ‘unofficial’ estimations of TCs by the ODPM. Nevertheless we call these ‘official’.

## 2. TCs and the Wider Urban System

One can draw on two main bodies of analysis, both trying to explain where TCs are and why they are important: central place theory (CPT) and gravitational theory. In the case of CPT, economists go first to Lösch (1940), although geographers might prefer the slightly earlier contribution of Christaller (1933). But both analyse essentially the same problem: Why does an urban system emerge and would emerge on even a flat and homogeneous plain? The essential mechanism is the tension between economies of scale and the costs of distance combined with the fact that some producers—farmers—are tied to the land and consume land in their production. Imagine a flat, fertile, and homogeneous plain with farmsteads dispersed over it. Over time some production gets concentrated in space because of economies of scale: so instead of all farmers brewing their own beer, for example, a brewery emerges serving the surrounding farms. The more important are economies of scale in any activity, the fewer will be the centres which end up producing that good other things equal. Similarly, the more significant are transport costs for any activity, the more centres will produce that good other things equal. So, we end up with a settlement pattern which has lots of brickworks and pubs but very few centres producing pharmaceuticals. The result is a hierarchy of places.

Translating this to the context of retail, we can think of the hierarchy of shopping centres. Many small places will offer convenience stores, but specialized fashion or department stores will be concentrated in a smaller number of larger shopping centres. In retail, as with other economic activities, there are economies of scale and a threshold market size necessary to support the activity. Rolls Royce dealers or bespoke tailors require large catchment areas (market sizes) to support them, so they are concentrated in fewer larger centres. If transport costs fall or the necessary minimum market size increases (the growth of Internet shopping may have increased the necessary minimum market size to support record or bookshops, e.g.), then there will tend to be an increase in concentration of retail in the larger centres: so the distribution of the 'hierarchy of shopping centres' will become more skewed.

CPT is a theory of a system of cities, of an urban hierarchy, and translates directly into a theory of a system of shopping centres. Some authors (Fujita et al. 1999) argue that CPT does not have testable assumptions and so should be only be considered as a descriptive theory. This argument is contested by researchers such as Denike & Parr (1970) who show there can be strict microfoundations for Christaller's model. In a similar vein, Dicken and Lloyd (1990) discuss testable hypotheses of the theory: in particular on the 'desire lines' (consumers' travel patterns or 'flows') within the hierarchy. Low-order goods (bread) generate short-distance and abundant 'desire lines' within a fine grid of central places, and high-order goods (furniture or cars) generate long-distance and fewer 'desire lines' within a coarse grid of central places.

We can think of CPT, therefore, as providing a theory of the system and hierarchy of shopping centres, but there is also a body of work which focuses on consumers' choices of where to shop and so on 'shopping trips'. As early as 1930, Reilly explored the location of retail (Reilly 1929, 1931). He presented a 'law of gravitation': areas of greater population ('mass') will generate more purchases in their centre, but their attraction will decay with the square of distance to any consumer or shopper. This theory was extended and refined by Huff (1963; 1964) taking as his inspiration, Newton's Law of Universal Gravitation. He described in a simple and powerful way the interactions between cities on a plain with dispersed population. This not only accounts for the length of shopping trips, increasing with the 'pull' of the shopping centre, the infrequency of that type of purchase, or a reduction in travel costs but also an emerging hierarchy of shopping centres of different sizes (Klaesson & Öner 2014).

Both these theories of cities and shopping trips can also be theories for TCs. Both can play a role in assessing the location, size, and distribution of TCs. In this article, we use an eclectic theory that draws heavily on both CPT and the gravity model approach. Specifically, we follow an econometric forecasting model initiated recently by Thurstain-Goodwin and Unwin (2000). We try to predict given TCs' locations, sizes, and distribution in one region using many variables, including proxies for 'mass' (population and area of retail as generally used in gravitational models) and 'desire lines' and hierarchies (drawing on CPT). Then, after verifying that there is a good fit, we predict the size of TCs in another region using the coefficients found in the first step.

## 3. What Is a TC and How Should It Be Identified?

As noted in the introduction, identifying the exact boundaries of TCs is a more challenging question to answer than it appears at first sight. TCs are not definite entities. They might not be located at the geometric or geographical centre of a city, and they might have fuzzy or indeterminate borders. The 'ideal' TC is not a point but is represented by a space of significant dimension. As the Oxford English Dictionary (OED) defines it: 'the central part or main business and commercial area of a town'. In general conversation, people might understand a TC to be the focal point of a city where main roads converge and people congregate. Historically the town or city centre was a place where citizens met or gathered: the place of the Italians' *passegiata*. Another function of a TC, captured in the OED definition, is as a space where jobs are concentrated, a shared workplace for people who live more spatially dispersed, and a centralized destination (workplace) for decentralized origins (households). Firms locate in TCs to be able to draw on a wider pool of labour. So, people commute to work in TCs. And the third main function of TCs is as a commercial hub, the space where people shop. 'High Streets' and market places are located in TCs.

But the space that represents a TC not only need not be at the geometric centre of a city, it does not have a unique shape. It would only be like that in a location that is constructed according to a rigidly imposed, utopian planning scheme, where all the uses and functions identified would be neatly and exclusive concentrated in only the TC, and TCs would have some uniform shape. Real TCs, in real cities, are much more messy and diverse, sometimes two or three blocks in the centre of a small town and sometimes very extensive. For example, Central London's DCLG 'designated' TC extends over 44 sq km, centred around Trafalgar Square, and includes many retail sub-centres, areas focused on business, and other specialized areas such as 'theatre land' or entertainment zones with a concentration of restaurants and nightlife. The diversity of real TCs certainly adds to choice and likely generates greater productivity and welfare. Left to choose for themselves, businesses and individuals will usually find superior locations to those decided on by urban planners, although there are significant qualifications resulting from externalities in land use that individualistic decision makers will tend to ignore.

If we are to reliably identify TC areas, then we ought to give due weight to the location of all the main functions discussed above to identify the location, size, and shape of the TC. All three aspects of TCs tend to be problematic theoretically and empirically. Centres do not need to be at the centroid of the city or some set of central jurisdictions. The observed shapes of TCs are motley and uneven. Size is also contentious. Empirically, in this article, we try to predict radiuses using a model with over 65 explanatory variables that capture all the multiple dimensions of 'town centredness'.

Attempts to provide operational definitions of TCs in Britain have been lead historically by what is now the DCLG (Thurstain-Goodwin and Unwin (2000); ODPM and CASA (2002); ODPM (2004); and more recently Dolega et al. (2016)). ODPM and CASA (2002) start by discussing a TC definition that depends on the perspective of a particular stakeholder. For instance, a taxi driver would have a different definition of a TC to a planner. For the taxi driver, the areas with the highest footfall can be determinants, while for a planner, the future evolution of the area might be a priority. Moreover, ODPM and CASA (2002) make the definition of TCs relative to other features of a city, creating an open approach from which they can build their model to define TCs.

The result is that their TCs are necessarily diverse. For some TCs the priority would be 'a retail core, and office centre and an area of high building density', while for others, 'a concentration of visitor attractions and associated retail outlets' would be the focus (Thurstain-Goodwin and Unwin 2000). What is meant by this is that it is essential to include multiple dimensions and functions, not just focus on one dimension of 'town centredness'. This implies that TCs are 'indeterminate objects' with fuzzy borders, extremely difficult to define and agree upon. We can add that an operational definition should be implemented with consistency over an entire set of cities because the identification of a TC remains problematic. For example, Wolverhampton's TC has a distinct ring road—some emergency services use it as a boundary, but administrative boundaries have been set in a much more extensive area reflecting a longer-term strategic vision of how the TC should evolve (ODPM and CASA 2002).

Typically, humans can easily detect an outlier, but not as easily notice when observations are clustered (Everitt and Hothorn 2011). Estimating *kernel density functions* can help identify clusters of 'objects'. These generate surfaces similar to mountainous terrain. This is called 'smoothing' and permits discrete and clustered data to be transformed into these mountain ranges. The kernel counts the number of observations in a given two-coordinate space as a histogram would, but it uses the number of observations to amplify a pulse function (rectangular, triangular, or normal most commonly) (Everitt and Hothorn 2011). Thus, waves effectively transform the discrete information of the numbers and intensities of the points into peaks and valleys. The key parameter is the bandwidth, which can be adjusted (Everitt and Hothorn 2011).

A very small bandwidth creates a single point to be counted independently, resulting in a spiky, disaggregated graph. An even smaller bandwidth provokes equal-sized extra-large pulse functions independent of each other if the observations are not located in exactly the same place. A very high bandwidth includes all points in a uniform one-shaped tiny image equal to the generating pulse kernel. Figure A1 (modified from Everitt and Hothorn 2011) shows an example of a one-dimensional normal kernel function for extremely low, low, optimal, and extremely high bandwidths. So, to be useful a researcher estimating kernel density functions needs to find a Goldilocks bandwidth neither too high nor too low. Many techniques have as a result been elaborated for finding such appropriate bandwidths. Then comes the next vital step: slicing the surfaces to get the curves or contour maps which are much easier to interpret. Thus, clustering can be detected by higher mountains, and areas, where data points are scarce, can be detected by lower ones.

Thurstain-Goodwin and Unwin (2000) define an index of intensity of 'town centredness' using the dimensions of property, economy, diversity, and visitor attractiveness. Because the categories are different in units, they employ a $z$-score normalization. The model is populated by points at the Unit Post Code (UPC) level (full postcodes), shaping town centredness as a mass function that is sliced for visualization. The intensity of the functions helps to delimit the border of the TCs, the visualization of which is the point of the study. The ODPM reports (ODPM and CASA 2002; ODPM 2004) are based on this methodology.

A catchment area is an area that draws in some group—customers or workers, for example. A gravity model adds some forces of attraction and repulsion. Gravity models are simple but can be empirically well-behaved and make good predictions. In the case of a retail centre, gravity models typically use square footage of retail space as a measure of size and travel time between retail centres for distance. The so-called 'Huff model' (Huff 1963) uses square footage as a directly proportional proxy of the number of products a consumer would find in each shopping centre and time as an inversely proportional proxy of the cost (including opportunity costs) of travelling to the given retail centres. Then, the more products there are and the greater quantity of a given product that is sold—represented by the square footage dedicated to a given kind of product—the greater the probability of visiting the given retail centre. And the lower the cost—measured as time—the greater the probability of visiting a given retail centre. The model has in the numerator the linear probability of the consumer choosing the retail good of a given type and in the denominator the sum of the linear probabilities of choosing all types of retail goods.

The Liverpool group, Dolega et al. (2016), discusses a method of defining TCs based on catchment areas. In summary, their method consists of replicating a catchment area for multiple stores. They use the Huff-model (Huff 1963, 1964, 2003) mentioned above. In this the probability, $P_{ij}$, that a consumer located at $i$ chooses to shop at retail centre $j$ is:

$$P_{ij} = \frac{A_j^{\alpha} D_{ij}^{-\beta}}{\sum_{j=1}^{n} A_j^{\alpha} D_{ij}^{-\beta}},$$

where:

$A_j$ is a measure of attractiveness of retail centre $j$, as square footage.
$D_{ij}$ is the distance from location $i$ to shop $j$.
$\alpha$ is the attractiveness parameter to be estimated.
$\beta$ is the distance decay parameter to be estimated.

Until recently the estimation of these parameters did not have known properties of large samples. Huff (2003) suggests it is necessary to explore alternative models similar to those presented in this article. In addition, Dolega et al. (2016) suggest that calibration at a national level would be superior to a local or subnational one. We also include a national-level estimation in our model.

The approach we take is more pragmatic and, in spirit, closer to Thurstain-Goodwin and Unwin (2000). We take the extent of the DCLG-defined TCs (their area-imputed radius) as 'true' on average and collate a long list of explanatory factors that we believe correlate with TC activities and characteristics to predict the TC radius. Then, having satisfied ourselves that the method provides sufficiently high goodness of fit, we use the estimated coefficients from this prediction to extrapolate out-of-sample and apply the coefficients to a different set of locations. Details of the data used for the estimates and the details of the method are explained more fully in the next two sections.

## 4. The Existing TC Data for England and Wales

As explained above, the first step of our methodology relies on the use of a given set of TC locations that we believe are reasonable approximations, as accurate a set of measures as is available: those identified by DCLG for England and Wales and as defined for 2000. Thurstain-Goodwin and Unwin (2000) and ODPM and CASA (2002) set out a methodology to identify what they call ATCAs, generalized to all locations in England and Wales in ODPM (2004). In the 2000 data, there are 1029 ATCAs, and additionally, within these ATCAs there are 46 RCs, giving a total of 1075 TCs for England and Wales.

The ATCAs are defined areas containing concentrations of 'town centre activity' aiming to be consistent with the theoretical basis summarized in Section 2. Both the hierarchy and the mass of TC activity are taken into consideration by the list of variables chosen to represent the point information with a kernel function. These 3D surfaces with heights reflecting TC activity are then sliced to form contour maps or level curves that represent locations with the same degree of TC activity. For instance, the concentration of employment is a direct measure of the mass of TC activity in gravity theory. At the same time, the postcode centrality structure is a direct measure of the CPT hierarchy.

The ATCAs were first constructed in a so-called Feasibility Study (DETR 1998)[6] using information on seven variables or elements: turnover, activities and facilities, pedestrian gateways, diversity, lack of resident population, intensity of use, and visitor attractions. In the follow-up London Pilot Study (ODPM and CASA 2002), these components were reduced to just three: economy, diversity, and property/intensity of use. Economy includes activities frequently found in TCs, such as retailing (convenience, comparison, and service retail); commercial offices; public administration; restaurant and licenced premises; arts, culture and entertainment; hotels; and public transport. This calculation implies the use of a set of very detailed values on variables reflecting employment (economy and diversity) and floor space (property), with a slightly less important use of turnover. The Office for National Statistics (ONS) contributed to the Inter-Departmental Business Register on employment and turnover for individual businesses, while the Valuation Office Agency—VOA—supplied an extensive commercial and industrial property floor space database.

The model identifies concentrations of the type of activities and patterns of property likely to be found in TCs where there are high levels of employment in economic activities common to TCs (including retail, offices, and leisure activities), a diversity of these activities, and a high density of office and retail floor space. Estimates are mapped at the detailed unit-level postcode to produce a surface of economic activity. Cutting through the peaks in the activity at a prescribed level for the whole of England and Wales gives the ATCA boundaries. Intuitively, combining employment and retail floor space data, a 3D data surface was constructed for different locations in England and Wales where the tallest peaks identified the largest concentrations of retail activity. Then, contours were drawn around these peaks, and the resulting areas were identified as ATCAs. In a second step, the data were cross-validated using external sources to make sure they corresponded to the main centres of activity in England and Wales.

Even if the ODPM/DCLG ATCAs are not intended to be operational for robust policy evaluation, since they correspond to revealed TC space and not planners' TCs as used for purposes of policy, they are the best definitions available to us, and their identification is based on high-quality data for very small geographical units. However, for the purposes of

6 Department of the Environment, Transport, and the Regions.

the evaluation (Cheshire et al. 2017), there exists an important limitation. This critical limitation is that these TCs are not defined for Scotland, and to evaluate the impact of TCFP, one needs to be able to compare developments in TCs in England and Wales, where the policy was strictly applied, to those in Scotland, where it was not. At the same time, we cannot replicate the exact methodology of ODPM/DCLG using data for Scotland because either these data are not readily available to us (e.g. the postcode-level information on different activities) or they do not exist for Scotland (e.g. the VOA data). Given these reasons, we opted to exploit the information on the size of the TCs that we can derive from the England and Wales set in the DCLG data, and combine it with a very rich data set on small geography explanatory factors (including socio-economic and topological features) that can successfully explain the variation in TC space we observe in the data.

## 5. Identifying TC Space for All Locations in Great Britain: Methodology

We combine data at small geographical scales from multiple sources to predict the extent of TCs for the whole of Great Britain. The main aim behind our methodology is to find a way to replicate 'as close as possible' the TC definitions available for England and Wales (ODPM 2004) and to be able to apply it to obtain TC boundaries in all cities in Britain. There are seven steps in our process:

1. **Select DCLG 2000 TC sample (DCLG TCs):** We start the process by exploring the DCLG list of TCs for England and Wales for the year 2000. From their observed surfaces we find the radius representing all the TCs as circular.[7] Then we select the samples for the regressions in Step 4. Of the 1075 TCs (1029 ATCAs and 46 RCs), we select two main samples: (i) all ATCAs; (ii) ATCAs and, for Central and West London, the RCs. From these samples, we drop the TCs which we consider cannot be used in the estimations.[8] To identify these, we use the information from the National Survey of Local Shopping Patterns (NSLSP) on the location of (grocery) shops in 1998 (more details are provided below).[9] The final samples

7 Of course, not all TCs need to be circular; although in the DCLG sample, they mostly are. Circles are one of the most efficient shapes to serve an area. In the case of TCs, because they occupy only a small fraction of the overall UK landscape, there is no need to impose a more efficient shape, such as a hexagon, to 'fill up space'.

8 As we use variables defined over 1 km of the centroid of the TC, those which do not have values defined within that radius were dropped from the sample.

9 The NSLSP is a yearly survey run by CBRE covering over a million households in the UK. Each sampled household is asked about their socio-economic characteristics, where they live and where they undertake their main shopping for a series of goods (groceries and household white goods). The data we obtained correspond to the grocery shopping locations and were aggregated spatially. It consists of an origin (postal sector)-destination (store) matrix of shopping trips. Postal sector areas are aggregations of postcodes and correspond to small areas (there are 12,000 in Britain). For the purposes of this article we used the shopping destination data to obtain a list of main (grocery) stores identified in 1998 as a grocery shopping destination in the NSLSP data. We can use this to infer how relatively important a TC is. This is illustrated in Figure 2. In addition, we used another CBRE-supplied data set on the location of retail units called RETLOC (RETail LOCations) in the main text, which includes information about all grocery stores and not only on those identified by the sample of households in the NSLSP as their grocery shopping destinations.

have between 810 and 950 TCs located in England and Wales. The mean radius of these is slightly less than 250 m. We then create centroids from the shapefiles of these TCs.

2. **Identify alternative TC locations for all Great Britain (OSC TCs):** We define an alternative list of TC centroid candidates using the towns and cities information in the OS Gazetteer towns and cities. Initially, there are 1315 towns and cities in Great Britain as a whole. As in the case of the DCLG TCs, the list is further trimmed when we combined it with the spatial data around the centroids. The exact location of some of these town and city centroids was 'relocated' by looking at where popular map navigation tools (such as Open Street Map or Google Maps) located the city centroid.
3. **Collection of data around the centroids of the DCLG and OSC TCs:** We collect abundant information at very small geographical scales (the largest is the Output Area and the smallest is postcode units) for the areas around the centroids of the DCLG and OSC TCs. The main results (presented in Section 6) use information around 1 km of the centroid, but we also calculated all the models using information around 2 and 3 km.[10] We believe that these long lists of socio-economic and topological features around 1 km of the centroid are sufficient satisfactorily to predict the extent of TC space around these centroids (remember the average DCLG TC radius is around 250 m). We obtained information on multiple variables (over 100) and 66 were used for the regressions of Step 4. The list of variables and their data sources appears in Table 1 (and in detail in Table A2).
4. **Estimation of the factors determining the extent of TC space:** For the DCLG TC samples selected in Step 1, we estimated several models where we explained the (log) radius of the TC as a function of the large set of explanatory variables around 1 km of the centroid of the TC. Inspired by the original ODPM models (2004), we use explanatory variables related to different town centredness dimensions (shop density and location, employment density and diversity, local amenities, socio-economic characteristics of the resident and working populations, infrastructure endowments, geographical location, physical barriers, etc.). The results of these regressions are shown in Table 2 and discussed in the next section. The majority of estimates are significantly different from 0, and the models have high goodness-of-fit statistics ($R^2$ between 0.78 and 0.88).
5. **Validation of the results (within DCLG sample):** The first step to validate our results is to check if the predictions correlate with the actual values for the in-sample. We use the coefficients estimated in Step 4 to predict the (log) radius of the DCLG TCs, both for the whole sample (1001) and for the samples used in each of the models estimated (referred as sub-samples in the tables). We both summarize and correlate the actual and predicted radius (and derived area) and use this to check the internal validity of the methodology. The results are shown in Tables 3 (and Table A3) and 4 and are discussed in the next section. They show that the statistical moments and the correlation between the actual and predicted (log) radius and area of the TCs are reasonably similar/high.
6. **Application of the model to predict TC space around the OSC TCs:** The results from Step 5 give us sufficient confidence that the models are satisfactorily accurate in their prediction of the extent of TCs for different values of the explanatory factors. We, therefore, proceed to apply the estimated betas from Step 4 to the 'out-of-sample' list of

10 These models had less predictive power, so we favoured the ones using 1 km. The average radius of a TC in England and Wales is 250 m, so we expect that values of the variables beyond 1 km of the TC centroid will have little power in predicting the extent of TC space. In fact, the heterogeneity of values beyond 1 km reduced the predictive power of the models.

**Table 1.** List of explanatory variables included in our model

| Variable | Data sources |
|---|---|
| Number of shoppers, shops, and location of these | *CBRE: RETLOC and NSLSP data in 1998* |
| Number and employment in sector 52 (retail) | *Annual Business Inquiry (ABI) accessed via NOMIS* |
| Population, residential employment, workplace-based employment, share of occupations (in workplace employment) of different levels (employment diversity), residential socio-economic characteristics (age structure, unemployment rates, labour market, commuting patterns) | *Census, 2001* |
| Transport infrastructure (roads, rail, buses) | *Ordnance Survey (OS) Strategi 2009, Open Street Map* |
| Cultural amenities (libraries, museums, art galleries, theatres, cinemas), consumption amenities (bars, restaurants), and historical amenities (landmarks, tourist info, local government) | *OS Strategi, OS Points of Interest (POI)* |
| Postcode centrality structure (sectors, districts, and towns) | *ONS National Statistics Postcode Directory (NSPD) and Wikipedia* |
| Geographical location (distance to the coast, river, rail, town hall, natural park/woodland) | *OS Strategi, POI, Wikipedia* |
| Topological features: terrain elevation (m) and slope (degrees) | *OS Panorama 50x50* |
| Nightlight brightness intensity (96–97 average) | *NOAA-NGDC*[11] |
| Postcode centrality structure (sectors, districts and towns) | *NSPD* |

*Notes*: List of the abbreviations used and more details on the variables provided in Tables A1 and A2.

OSC TCs and calculate the predicted (log) radius and area for these locations. This generates a set of estimated surrogate TC shapefiles to cover all the TCs of Great Britain. We can compare the predicted radius for the two sets of TCs (DCLG and OSC) for the sample which is available in both data sets (e.g. England and Wales together and England and Wales separately). This is done in the first two rows of Table 5 and shows that the values of the DCLG sample and our OSC TC predicted values are similar.

7. **Comparison of socio-economic variables within the DCLG and OSC TCs**: The DCLG TCs and our predicted OSC TCs differ in two dimensions: their particular size for a given set of explanatory factors (which we fit in Step 4) and their specific location. The precise places where the OSC and DCLG centroids are located can differ, and, in particular, there is no comparison group for Scotland. To overcome this, in Step 7 we calculate several socio-economic descriptive statistics (population, number of addresses, number of shoppers, etc.) within the boundaries (or a small distance of them) of the two sets of TCs. The summary statistics for these are shown in the remaining rows of Table 5. These allow us to check whether, even when located at slightly in different places, the underlying economic factors within TC boundaries are comparable in the two samples and, additionally, to explore how different the Scottish TCs are compared to those in England and Wales.

11 National Oceanic and Atmospheric Administration; National Geophysical Data Center, non-censored version.

**Table 2.** TC extent prediction model results

| DCLG 2000 TC sample | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Type of TC | ATCA | ATCA | ATCA | ATCA | ATCA | ATCA |
| Fuzzy TC border | 10 m | 100 m | 500 m | 10 m | 100 m | 500 m |
| Central and West London | ATCA | ATCA | ATCA | RC | RC | RC |
| Number of observations | 812 | 870 | 931 | 824 | 882 | 949 |
| Adjusted $R^2$ | 0.878 | 0.868 | 0.810 | 0.841 | 0.834 | 0.779 |
| $R^2$ | 0.888 | 0.878 | 0.824 | 0.854 | 0.846 | 0.795 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Log of grocery shoppers | 0.020*** | 0.024*** | 0.031*** | 0.021** | 0.025*** | 0.032*** |
| (NSLSP98), 1 km of centroid | (0.007) | (0.007) | (0.009) | [0.009] | [0.009] | [0.009] |
| Log number of shops | 0.024*** | 0.026*** | 0.021*** | 0.022*** | 0.024*** | 0.018*** |
| (RETLOC/NSLSP98), 1 km of centroid | (0.003) | (0.003) | (0.004) | (0.003) | (0.003) | (0.004) |
| Log average distance to shops | −0.441*** | −0.401*** | −0.684*** | −0.435*** | −0.396*** | −0.631*** |
| (RETLOC98), 1 km of centroid | (0.123) | (0.123) | (0.144) | (0.131) | (0.133) | (0.146) |
| Log average distance to shops | −0.177*** | −0.161*** | −0.283*** | −0.175*** | −0.158*** | −0.261*** |
| (RETLOC98), 1 km of centroid$^2$ | (0.054) | (0.055) | (0.064) | (0.059) | (0.060) | (0.065) |
| Log average distance to shops | −0.021*** | −0.019*** | −0.034*** | −0.021*** | −0.019** | −0.031*** |
| (RETLOC98), 1 km of centroid$^3$ | (0.007) | (0.007) | (0.008) | (0.008) | (0.008) | (0.008) |
| Log average distance to shops | −0.051 | −0.080 | −0.149 | −0.046 | −0.090 | −0.155 |
| (NSLSP98), 1 km of centroid | (0.095) | (0.095) | (0.115) | (0.093) | (0.096) | (0.114) |
| Log average distance to shops | −0.015 | −0.021 | −0.045 | −0.011 | −0.024 | −0.046 |
| (NSLSP98), 1 km of centroid$^2$ | (0.043) | (0.043) | (0.052) | (0.043) | (0.044) | (0.053) |
| Log average distance to shops | −0.002 | −0.002 | −0.004 | −0.002 | −0.003 | −0.004 |
| (NSLSP98), 1 km of centroid$^3$ | (0.006) | (0.006) | (0.007) | (0.006) | (0.006) | (0.007) |
| Log distance to closest shop | 0.054*** | 0.046*** | 0.013 | 0.059*** | 0.050*** | 0.018* |
| (NSLSP98), 1 km of centroid | (0.010) | (0.010) | (0.011) | (0.011) | (0.010) | (0.011) |

(continued)

**Table 2.** Continued

| DCLG 2000 TC sample | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Number of units retail sector | 0.001*** | 0.001*** | 0.001*** | 0.001*** | 0.001*** | 0.002*** |
| (ABI98), 1 km of centroid | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Number of units retail sector | −0.000* | −0.000** | −0.000 | −0.000*** | −0.000*** | −0.000*** |
| (ABI98), 1 km of centroid$^2$ | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Employment retail sector | −0.000 | −0.000 | −0.000 | −0.000 | −0.000* | −0.000** |
| (ABI98), 1 km of centroid | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Employment retail sector | 0.000 | 0.000* | 0.000 | 0.000** | 0.000*** | 0.000*** |
| (ABI98), 1 km of centroid$^2$ | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Log density of total population | 0.362 | 0.211 | 0.457* | 0.323 | 0.180 | 0.325 |
| (Census01), 1 km of centroid | (0.254) | (0.230) | (0.251) | (0.251) | (0.228) | (0.238) |
| Log density of total population | −0.024 | −0.015 | −0.032* | −0.020 | −0.012 | −0.022 |
| (Census01), 1 km of centroid$^2$ | (0.016) | (0.015) | (0.016) | (0.016) | (0.015) | (0.016) |
| Log workplace employment | 0.286*** | 0.284*** | 0.282*** | 0.281*** | 0.281*** | 0.279*** |
| (Census01), 1 km of centroid | (0.024) | (0.024) | (0.028) | (0.026) | (0.026) | (0.031) |
| Share wemple high occupations | −1.507*** | −1.647*** | −1.238** | −2.391*** | −2.444*** | −2.130*** |
| (Census01), 1 km of centroid | (0.481) | (0.465) | (0.573) | (0.505) | (0.494) | (0.553) |
| Share wemple medium occupation | −1.838*** | −1.784*** | −1.308** | −2.577*** | −2.434*** | −2.032*** |
| (Census01), 1 km of centroid | (0.474) | (0.457) | (0.576) | (0.522) | (0.507) | (0.578) |
| Share wemple low occupations | −2.292*** | −2.433*** | −2.104*** | −3.155*** | −3.201*** | −3.009*** |
| (Census01), 1 km of centroid | (0.490) | (0.473) | (0.581) | (0.515) | (0.503) | (0.559) |
| Average commuting distance | 0.013 | 0.012 | 0.012 | 0.018 | 0.017 | 0.018 |
| (Census01), 1 km of centroid | (0.009) | (0.008) | (0.009) | (0.011) | (0.011) | (0.011) |
| Average commuting distance | −0.000 | −0.000 | −0.000 | −0.000 | −0.000 | −0.000 |
| (Census01), 1 km of centroid$^2$ | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |

(continued)

**Table 2.** Continued

| DCLG 2000 TC sample | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Percentage commuters by foot/bike | 0.110 | 0.125 | 0.216 | 0.307** | 0.305** | 0.383** |
| (Census01), 1 km of centroid | (0.142) | (0.145) | (0.157) | (0.144) | (0.146) | (0.155) |
| Percentage commuters public trans | 0.018 | 0.083 | 0.107 | 0.294** | 0.344*** | 0.279** |
| (Census01), 1 km of centroid | (0.114) | (0.114) | (0.128) | (0.120) | (0.119) | (0.130) |
| Average household size | 0.100* | 0.030 | 0.042 | 0.141** | 0.079 | 0.067 |
| (Census01), 1 km of centroid | (0.057) | (0.064) | (0.071) | (0.065) | (0.068) | (0.076) |
| Average age of resident population | 0.019*** | 0.013* | 0.009 | 0.028*** | 0.022*** | 0.018** |
| (Census01), 1 km of centroid | (0.006) | (0.007) | (0.007) | (0.007) | (0.007) | (0.007) |
| Percentage of population aged 18–44 years | 0.297 | 0.184 | −0.064 | 0.733** | 0.626* | 0.571 |
| (Census01), 1 km of centroid | (0.273) | (0.280) | (0.333) | (0.325) | (0.330) | (0.368) |
| Total unemployment rate | 0.819*** | 0.815*** | 0.853*** | 1.178*** | 1.160*** | 1.298*** |
| (Census01), 1 km of centroid | (0.285) | (0.290) | (0.326) | (0.293) | (0.295) | (0.341) |
| Share of students in population | −0.933*** | −0.890*** | −0.663*** | −1.043*** | −0.983*** | −0.842*** |
| (Census01), 1 km of centroid | (0.183) | (0.180) | (0.204) | (0.233) | (0.222) | (0.247) |
| Share of retirees in population | −0.699 | −0.476 | −0.274 | −0.842* | −0.630 | −0.333 |
| (Census01), 1 km of centroid | (0.450) | (0.453) | (0.475) | (0.478) | (0.477) | (0.501) |
| Log of km of all roads | −0.005 | 0.004 | 0.007 | −0.001 | 0.004 | 0.013 |
| (Strategi 2009), 1 km of centroid | (0.022) | (0.022) | (0.025) | (0.024) | (0.024) | (0.025) |
| Number of bus stations | −0.001 | −0.000 | −0.001 | 0.005 | 0.004 | 0.004 |
| (POI 2015), 1 km of centroid | (0.008) | (0.008) | (0.009) | (0.008) | (0.008) | (0.010) |
| Number of tube/tram stations | 0.059** | 0.057** | 0.001 | −0.048** | −0.047** | −0.056*** |
| (POI 2015), 1 km of centroid | (0.023) | (0.025) | (0.026) | (0.022) | (0.021) | (0.019) |
| Number of rail stations | 0.003 | −0.000 | −0.012 | 0.000 | −0.002 | −0.010 |
| (POI 2015), 1 km of centroid | (0.010) | (0.010) | (0.014) | (0.015) | (0.015) | (0.015) |

(continued)

**Table 2.** Continued

| DCLG 2000 TC sample | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Number of libraries | −0.014 | −0.011 | −0.032** | −0.028** | −0.026** | −0.045*** |
| (POI 2015), 1 km of centroid | (0.010) | (0.011) | (0.015) | (0.012) | (0.012) | (0.014) |
| Number of museums | 0.004 | −0.004 | −0.000 | 0.021* | 0.011 | 0.010 |
| (Strategi 2009), 1 km of centroid | (0.012) | (0.012) | (0.013) | (0.013) | (0.013) | (0.014) |
| Number of art galleries | −0.010* | −0.009* | −0.010* | −0.005 | −0.004 | −0.009 |
| (POI 2015), 1 km of centroid | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.006) |
| Number of cinemas and theatres | 0.017*** | 0.016*** | 0.017** | 0.018** | 0.017** | 0.019** |
| (POI 2015), 1 km of centroid | (0.006) | (0.006) | (0.008) | (0.008) | (0.008) | (0.008) |
| Number of discos and nightclubs | 0.004 | 0.003 | 0.002 | 0.022*** | 0.021*** | 0.021*** |
| (POI 2015), 1 km of centroid | (0.004) | (0.004) | (0.005) | (0.005) | (0.005) | (0.005) |
| Number of landmarks | 0.027* | 0.020 | 0.005 | 0.035** | 0.030** | 0.019 |
| (Strategi 2009), 1 km of centroid | (0.015) | (0.015) | (0.021) | (0.014) | (0.014) | (0.021) |
| Number of cafes, restaurants and pubs | 0.001* | 0.001** | 0.001** | −0.001** | −0.001* | −0.001** |
| (POI 2015), 1 km of centroid | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Number of B&B, hotels and motels | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| (POI 2015), 1 km of centroid | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Number of youth hostels | −0.007 | −0.008 | −0.006 | −0.012 | −0.013* | −0.008 |
| (POI 2015), 1 km of centroid | (0.007) | (0.007) | (0.008) | (0.008) | (0.008) | (0.008) |
| Number of local government sites | 0.015*** | 0.016*** | 0.013 | 0.012 | 0.013* | 0.009 |
| (POI 2015), 1 km of centroid | (0.005) | (0.005) | (0.010) | (0.007) | (0.007) | (0.011) |
| Number of tourist info offices | 0.012 | 0.015 | 0.023* | 0.012 | 0.014 | 0.025* |
| (Strategi 2009), 1 km of centroid | (0.012) | (0.012) | (0.013) | (0.013) | (0.012) | (0.014) |
| Number of visitor centres | −0.002 | −0.001 | 0.004 | −0.005 | −0.004 | 0.006 |
| (Strategi 2009), 1 km of centroid | (0.007) | (0.006) | (0.008) | (0.008) | (0.008) | (0.009) |

(continued)

**Table 2.** Continued

| DCLG 2000 TC sample | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Share of central addresses | −0.000 | 0.002 | −0.011 | 0.001 | 0.004 | −0.014 |
| (pcsect in pctown), 1 km of centroid | (0.031) | (0.030) | (0.031) | (0.031) | (0.029) | (0.030) |
| Share of central addresses | 0.034** | 0.038** | 0.030 | 0.034* | 0.038** | 0.028 |
| (pcdistr in pctown), 1 km of centroid | (0.017) | (0.017) | (0.019) | (0.019) | (0.019) | (0.020) |
| Log distance in km to first | −0.002 | −0.004 | −0.009 | −0.002 | −0.003 | −0.008 |
| postcode in the closest postal town | (0.005) | (0.005) | (0.006) | (0.005) | (0.005) | (0.005) |
| Log of distance to closest town hall | −0.003 | −0.003 | −0.001 | −0.005 | −0.005 | −0.004 |
| (Wikipedia), in km | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| Log of distance to closest rail or tube station | −0.002 | −0.000 | −0.010 | −0.006 | −0.004 | −0.011 |
| (POI 2015), in km | (0.006) | (0.006) | (0.008) | (0.008) | (0.008) | (0.008) |
| Log of distance to closest point in the coastline | −0.005 | −0.000 | 0.008 | −0.003 | 0.001 | 0.007 |
| (Strategi 2009), in km | (0.005) | (0.005) | (0.006) | (0.006) | (0.006) | (0.007) |
| Log of distance to closest river or lake | −0.008 | 0.000 | 0.008 | −0.013* | −0.005 | 0.002 |
| (Strategi 2009), in km | (0.006) | (0.006) | (0.006) | (0.007) | (0.007) | (0.007) |
| Log of distance to closest park or woodland | 0.001 | −0.003 | −0.003 | 0.001 | −0.004 | −0.003 |
| (Strategi 2009), in km | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) |
| Standard dev of elevation | 0.001 | 0.000 | 0.001 | 0.002 | 0.001 | 0.002 |
| (Panorama), 1 km of centroid | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Mean of elevation | −0.010* | −0.009* | −0.010* | −0.005 | −0.004 | −0.009 |
| (Panorama), 1 km of centroid | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.006) |
| Maximum of elevation | 0.002* | 0.001 | 0.002** | 0.001 | 0.001 | 0.002* |
| (Panorama), 1 km of centroid | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Range of elevation | −0.002 | −0.001 | −0.002** | −0.002 | −0.001 | −0.002** |
| (Panorama), 1 km of centroid | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |

(continued)

**Table 2.** Continued

| DCLG 2000 TC sample | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Standard dev of terrain slope | 0.016 | 0.018 | 0.022 | 0.015 | 0.017 | 0.017 |
| (Panorama), 1 km of centroid | (0.017) | (0.016) | (0.017) | (0.018) | (0.017) | (0.017) |
| Mean of terrain slope | 0.011 | 0.009 | 0.012 | 0.008 | 0.007 | 0.010 |
| (Panorama), 1 km of centroid | (0.011) | (0.011) | (0.011) | (0.012) | (0.012) | (0.012) |
| Maximum of terrain slope | 0.001 | −0.010 | 0.011 | −0.020 | −0.019 | 0.001 |
| (Panorama), 1 km of centroid | (0.127) | (0.124) | (0.122) | (0.131) | (0.129) | (0.127) |
| Range of terrain slope | −0.004 | 0.008 | −0.013 | 0.017 | 0.017 | −0.003 |
| (Panorama), 1 km of centroid | (0.127) | (0.124) | (0.122) | (0.131) | (0.129) | (0.127) |
| Standard dev of lights brightness | 0.005 | 0.003 | 0.006 | −0.004 | −0.004 | 0.000 |
| (NASA 96-97), 1 km of centroid | (0.005) | (0.005) | (0.007) | (0.005) | (0.005) | (0.006) |
| Mean of lights brightness | 0.000 | 0.001 | 0.001 | 0.000 | 0.001 | 0.001 |
| (NASA 96-97), 1 km of centroid | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.002) |
| Max of lights brightness | 0.001 | 0.000 | 0.000 | 0.000 | −0.000 | 0.000 |
| (NASA 96-97), 1 km of centroid | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Range of lights brightness | −0.002 | −0.001 | −0.003 | 0.001 | 0.001 | −0.001 |
| (NASA 96-97), 1 km of centroid | (0.002) | (0.002) | (0.003) | (0.002) | (0.002) | (0.002) |
| Sum of lights brightness | −0.000 | −0.000* | −0.000* | −0.000 | −0.000 | −0.000 |
| (NASA 96-97), 1 km of centroid | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Constant term (excluded: occupation | −4.977*** | −3.948*** | −5.621*** | −4.656*** | −3.798*** | −4.972*** |
| other and commuter by motor vehicle) | (1.138) | (1.063) | (1.252) | (1.143) | (1.078) | (1.193) |

*Notes*: Log stands for natural logarithm and km for kilometres. Details on the definition of the variables are given in Table A2.

**Table 3.** England and Wales TC radius/area (actual and predicted), and number of TCs, DCLG 2000 sample

| | TC derived radius in km | | | | | TC Area in square km | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Observed | Mean | Standard deviation | Minimum | Maximum | Observed | Mean | Standard deviation | Minimum | Maximum |
| *Observed value* | *1001* | *0.247* | *0.153* | *0.113* | *3.187* | *1001* | *0.265* | *1.047* | *0.040* | *31.905* |
| Average all samples | 1001 | 0.261 | 0.165 | 0.107 | 2.973 | 1001 | 0.308 | 0.994 | 0.036 | 28.018 |
| Sample 1 betas | 1001 | 0.273 | 0.196 | 0.104 | 3.095 | 1001 | 0.355 | 1.223 | 0.034 | 30.091 |
| Sample 2 betas | 1001 | 0.270 | 0.195 | 0.106 | 3.076 | 1001 | 0.350 | 1.203 | 0.035 | 29.718 |
| Sample 3 betas | 1001 | 0.260 | 0.178 | 0.105 | 3.051 | 1001 | 0.312 | 1.059 | 0.034 | 29.245 |
| Sample 4 betas | 1001 | 0.257 | 0.148 | 0.105 | 2.526 | 1001 | 0.276 | 0.718 | 0.034 | 20.040 |
| Sample 5 betas | 1001 | 0.255 | 0.153 | 0.107 | 2.711 | 1001 | 0.277 | 0.809 | 0.036 | 23.088 |
| Sample 6 betas | 1001 | 0.249 | 0.163 | 0.106 | 3.382 | 1001 | 0.278 | 1.184 | 0.035 | 35.929 |
| Sub-sample 1 | 812 | 0.256 | 0.157 | 0.104 | 3.095 | 812 | 0.283 | 1.090 | 0.034 | 30.091 |
| Sub-sample 2 | 870 | 0.250 | 0.153 | 0.106 | 3.076 | 870 | 0.270 | 1.044 | 0.035 | 29.718 |
| Sub-sample 3 | 931 | 0.243 | 0.147 | 0.111 | 3.051 | 931 | 0.254 | 0.996 | 0.039 | 29.245 |
| Sub-sample 4 | 824 | 0.251 | 0.120 | 0.105 | 1.040 | 824 | 0.243 | 0.306 | 0.034 | 3.397 |
| Sub-sample 5 | 882 | 0.246 | 0.118 | 0.107 | 1.070 | 882 | 0.233 | 0.302 | 0.036 | 3.600 |
| Sub-sample 6 | 949 | 0.239 | 0.112 | 0.108 | 1.030 | 949 | 0.218 | 0.281 | 0.036 | 3.336 |

*Notes*: The DCLG sample for which we have values of the variables around 1 km of the centroids corresponds to 1001 TCs in England and Wales. Rows Sample 1 betas–Sample 6 betas apply the coefficients estimated in Columns 1–6 of Table 2 to these 1001 locations. Sub-samples refer to applying the estimated coefficients of each column of Table 2 to the particular sample used in that estimation, for example, the number of TCs corresponds to the number of observations in each of these regressions.

**Table 4.** Correlation coefficients of real versus predicted values for radius and area, DCLG 2000 sample

| | All (1001) | | | Sub-sample | | |
|---|---|---|---|---|---|---|
| | *All* | *England* | *Wales* | *All* | *England* | *Wales* |
| TC radius in km | | | | | | |
| Average all samples | 0.755 | 0.757 | 0.866 | | | |
| Sample 1 | 0.722 | 0.718 | 0.912 | 0.970 | 0.970 | 0.975 |
| Sample 2 | 0.725 | 0.721 | 0.912 | 0.966 | 0.966 | 0.973 |
| Sample 3 | 0.803 | 0.800 | 0.923 | 0.948 | 0.948 | 0.967 |
| Sample 4 | 0.867 | 0.866 | 0.912 | 0.943 | 0.942 | 0.977 |
| Sample 5 | 0.870 | 0.869 | 0.913 | 0.940 | 0.938 | 0.973 |
| Sample 6 | 0.883 | 0.882 | 0.921 | 0.915 | 0.913 | 0.966 |
| TC area in km$^2$ | | | | | | |
| Average all samples | 0.928 | 0.928 | 0.865 | | | |
| Sample 1 | 0.810 | 0.810 | 0.858 | 0.994 | 0.994 | 0.974 |
| Sample 2 | 0.811 | 0.811 | 0.857 | 0.993 | 0.993 | 0.974 |
| Sample 3 | 0.914 | 0.914 | 0.873 | 0.988 | 0.988 | 0.958 |
| Sample 4 | 0.940 | 0.941 | 0.863 | 0.937 | 0.937 | 0.977 |
| Sample 5 | 0.950 | 0.950 | 0.862 | 0.928 | 0.927 | 0.975 |
| Sample 6 | 0.973 | 0.973 | 0.868 | 0.893 | 0.892 | 0.955 |
| **Sample sizes** | | | | | | |
| Average all samples | 1001 | 944 | 57 | | | |
| Sample 1 | 1001 | 944 | 57 | 812 | 768 | 44 |
| Sample 2 | 1001 | 944 | 57 | 870 | 820 | 50 |
| Sample 3 | 1001 | 944 | 57 | 931 | 876 | 55 |
| Sample 4 | 1001 | 944 | 57 | 824 | 780 | 44 |
| Sample 5 | 1001 | 944 | 57 | 882 | 832 | 50 |
| Sample 6 | 1001 | 944 | 57 | 949 | 894 | 55 |

To illustrate the logic behind our methodology, Figure 1 shows a flowchart depicting the seven steps explained above and the relationship between them. In the next section, we apply these steps to our data and discuss the results and the validation checks carried out.

## 6. Regression Results and Validity Tests

The first step of our methodology concerns the selection of the samples of the TC locations used in the estimations of the models that predict TC extent. The DCLG 2000 TC data set originally contained 1075 units. When we calculate the variables included in the estimation of Step 4, within 1 km of the centroid, a number of TCs are dropped from the sample because the values of some of these factors do not exist within that distance radius

**Table 5**. Predicted size and socio-economics for OSC and DCLG's TCs (500 fuzzy boundary tolerance)

| | England and Wales | | England | | Wales | | OSC TC areas | |
|---|---|---|---|---|---|---|---|---|
| Variable | OSC | DCLG | OSC | DCLG | OSC | DCLG | ALL GB | SCOT |
| Number of observations | *752* | *931* | *687* | *876* | *65* | *55* | *861* | *109* |
| Predicted TC radius in km | 0.231 | 0.246 | 0.236 | 0.248 | 0.187 | 0.210 | 0.222 | 0.161 |
| Predicted TC area in km$^2$ | 0.215 | 0.257 | 0.222 | 0.263 | 0.135 | 0.161 | 0.201 | 0.109 |
| NSLSP 1998 shops | 2.20 | 2.37 | 2.23 | 2.37 | 1.92 | 2.33 | 2.20 | 2.17 |
| NSLSP 1998 shoppers | 20,910.63 | 24,085.03 | 21,560.91 | 24,416.07 | 14,037.62 | 18,914.91 | 20,003.98 | 13,748.96 |
| NSPD address counts | 3183.63 | 4175.27 | 3291.85 | 4266.92 | 2039.77 | 2715.62 | 3044.47 | 2084.39 |
| NSPD small businesses counts | 419.88 | 479.08 | 433.57 | 487.45 | 275.25 | 345.84 | 398.52 | 251.14 |
| NSLSP 1998 shops per km$^2$ | 1.26 | 1.31 | 1.26 | 1.30 | 1.26 | 1.46 | 1.29 | 1.50 |
| NSLSP 1998 shoppers per km$^2$ | 11,387.03 | 12,928.03 | 11,651.57 | 13,032.77 | 8591.06 | 11,292.24 | 11,100.73 | 9125.53 |
| NSPD address counts per km$^2$ | 1762.30 | 2130.30 | 1806.99 | 2162.26 | 1289.95 | 1621.23 | 1716.77 | 1402.70 |
| NSPD small businesses counts per km$^2$ | 208.51 | 217.23 | 212.86 | 218.54 | 162.48 | 196.36 | 200.70 | 146.79 |
| Population | 6004.47 | 8192.70 | 6221.02 | 8389.54 | 3715.80 | 5158.30 | 5716.36 | 3728.64 |
| Residential employment | 2699.94 | 3708.18 | 2821.36 | 3817.85 | 1416.57 | 2017.62 | 2563.13 | 1619.28 |
| Workplace employment | 7438.53 | 8828.14 | 7792.53 | 9088.64 | 3697.03 | 4812.55 | 6986.72 | 3869.62 |
| Workplace high occupations | 3292.30 | 3972.25 | 3474.86 | 4113.57 | 1362.80 | 1793.74 | 3072.93 | 1559.44 |
| Share of work employment high occupations | 0.350 | 0.363 | 0.353 | 0.366 | 0.310 | 0.327 | 0.345 | 0.314 |
| Ratio work employment over population | 1.05 | 0.93 | 1.07 | 0.94 | 0.79 | 0.84 | 1.01 | 0.75 |
| Population per km$^2$ | 3718.12 | 4487.26 | 3785.34 | 4563.91 | 3007.71 | 3320.25 | 3634.28 | 3055.87 |
| Residential employment per km$^2$ | 1658.80 | 2019.03 | 1706.66 | 2066.52 | 1153.01 | 1295.91 | 1616.10 | 1321.48 |
| Work employment per km$^2$ | 3501.33 | 3366.00 | 3611.98 | 3417.46 | 2331.83 | 2582.36 | 3340.43 | 2230.32 |

*Notes*: The first three columns compare mean values of the variables for the TCs in the common sample countries (England and Wales). The last column provides the mean values of the variables for the whole GB sample and the areas in Scotland. TC stands for Town Centre, OSC corresponds to Ordnance Survey cities sample and DCLG to the ODPM/DCLG Town Centres; GB stands for Great Britain, km stands for kilometres, and NSLSP stands for National Survey of Local Shopping Patterns.

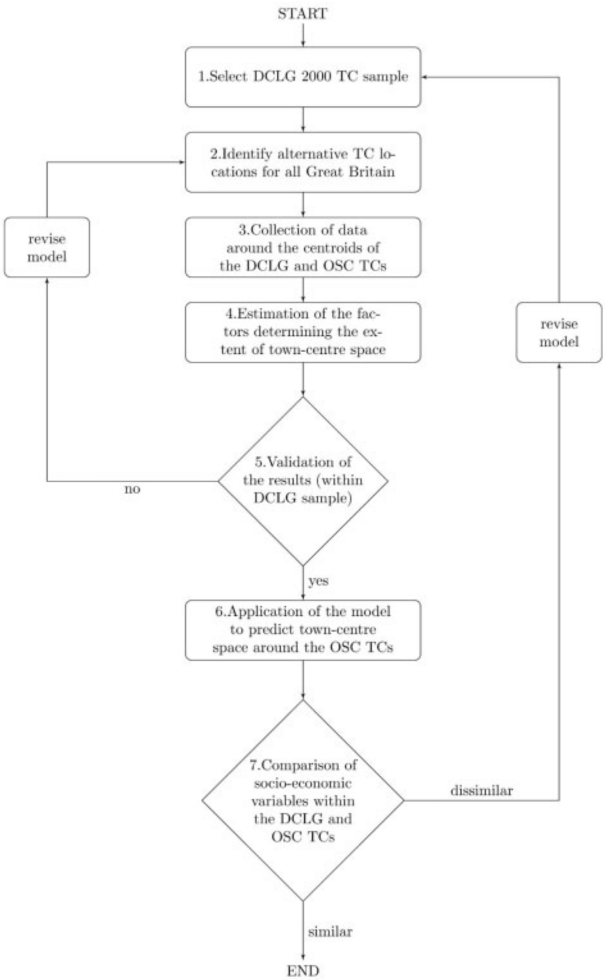


**Figure 1.** Flow chart illustrating the seven steps of our methodology.

(in particular, for the 1 km data set, we are left with 1001 TCs).[12] In addition, the two types of TC in the DCLG sample, ATCAs and RCs, overlap, so we do not want to use all of the polygons in our regressions. The samples used in the estimations presented in Table 2 differ in the set of DCLG TCs which are used. In Columns (1)–(3) of Table 2, we use all the ATCAs and none of the RCs. However, for Central and West London, the ATCAs are very large, and they mask the richness of small sub-centres (or towns) within them, as is depicted in Figure A2. In Columns (4)–(6) of Table 2, we use the ATCAs in all England and Wales, but for the Central and West London areas, we use the RCs (dark areas).[13]

12 The same occurs in the OSC sample—from 1315 we are left with 964 locations when we use the values of the explanatory variables

13 In all the maps the background geographic areas are the postal sectors.

Within each of these two samples (all ATCAs or ATCAs and London RCs), we introduce an additional criterion to select which TCs to include in the estimations of Step 4. Some of these TCs are certainly very small (25% of the ATCAs have an area of less than 0.08 sq km and an implied radius of less than 160 m). The NSLSP 1998 data allow us to map a set of approximately 4700 shops which consumers identify as their main grocery shopping destinations. Given the very large size of the NSLSP sample—more than 1 million households a year—it seems reasonable to identify TCs which do not contain any of these shops within a certain distance of their boundaries as 'less important'. This is illustrated in Figure 2: for areas around Manchester and Glasgow, we plot (tiny triangles) the NSLSP shops in 1998. We calculate, for both the DCLG and the OSC samples, the number of shops (and shoppers that choose those shops) within different distances of the TC boundary. We can choose an *ad hoc* threshold beyond which we consider the shop too far to be part of that TC.[14] A TC can have shops strictly inside its boundaries, within some allowed close distance of its boundary (fuzzy) or beyond an allowed distance of the boundary.[15] In the full results, we used six distance tolerance levels (fuzzy boundaries): 0 (at least one shop completely within the TCs), 10, 100, 250, 500 m, and 1 km. Without loss of generality, for the regression results provided in the article, we focus on 10, 100, and 500 m. The use of this restriction is what makes the sample size in Columns 1–6 differs from one another. It is worth noticing that the stricter we are with the criterion of at least one NSLSP shop in the TCs, the higher is the explanatory power of the models of Table 2. In Steps 2 and 6, we also use the fuzzy boundary criterion to select which OSC TCs are relevant in our final samples.

In Step 3, we select a large number of explanatory factors to predict the extent of TCs. We choose factors that we believe relate to TC activities. This step involves the collection of potentially relevant variables; GIS work to geographically match the data; and then choosing what variables to include in the final empirical model mainly on the basis of intuition and goodness of fit. This is akin to a forecasting and descriptive process, so we do not pay serious attention to multicollinearity but to the overall validity and explanatory power of the prediction models.

The specific list of variables used in the regressions of Step 4 is inspired by previous attempts in the construction of British TCs.[16] In particular, in the construction of the Index of Town Centredness discussed in the documents and papers that describe the construction

14 We consider the boundaries of the DCLG TCs to be subject to some level of measurement error. Therefore, being very strict about the location of NSLSP shops with respect to the TC boundaries would result in dropping many TCs from the samples. For this reason we adopt a flexible position and try using three different thresholds when selecting the TCs in the different estimating samples.

15 In the map for Manchester we observe all these cases: first inside the ATCA area of Eccles, there are four shops (tiny triangles), while the Trafford Centre has a nearby shop outside its ATCA area but probably inside both a 100 and 500m buffer of its boundary. Finally, Oldham Road, close to the Manchester metropolitan area, has no nearby shop, so it should be dropped from our sample if one of our many restrictions applies. In the map for Glasgow, there are no ATCA areas —because there are no 'official' areas defined for Scotland, only our predictions. These are shown as dark circles around Glasgow and Renfrew. Inside Glasgow's predicted TC, there are six shops (tiny triangles) and one nearby probably at a buffer distance of 10,100 and 500 m.

16 Most relevantly those of ODPM and CASA (2002), ODPM (2004) and Thurstain-Goodwin and Unwin (2000), but also Dolega et al. (2016) and Pavlis, Dolega and Singleton (2017).

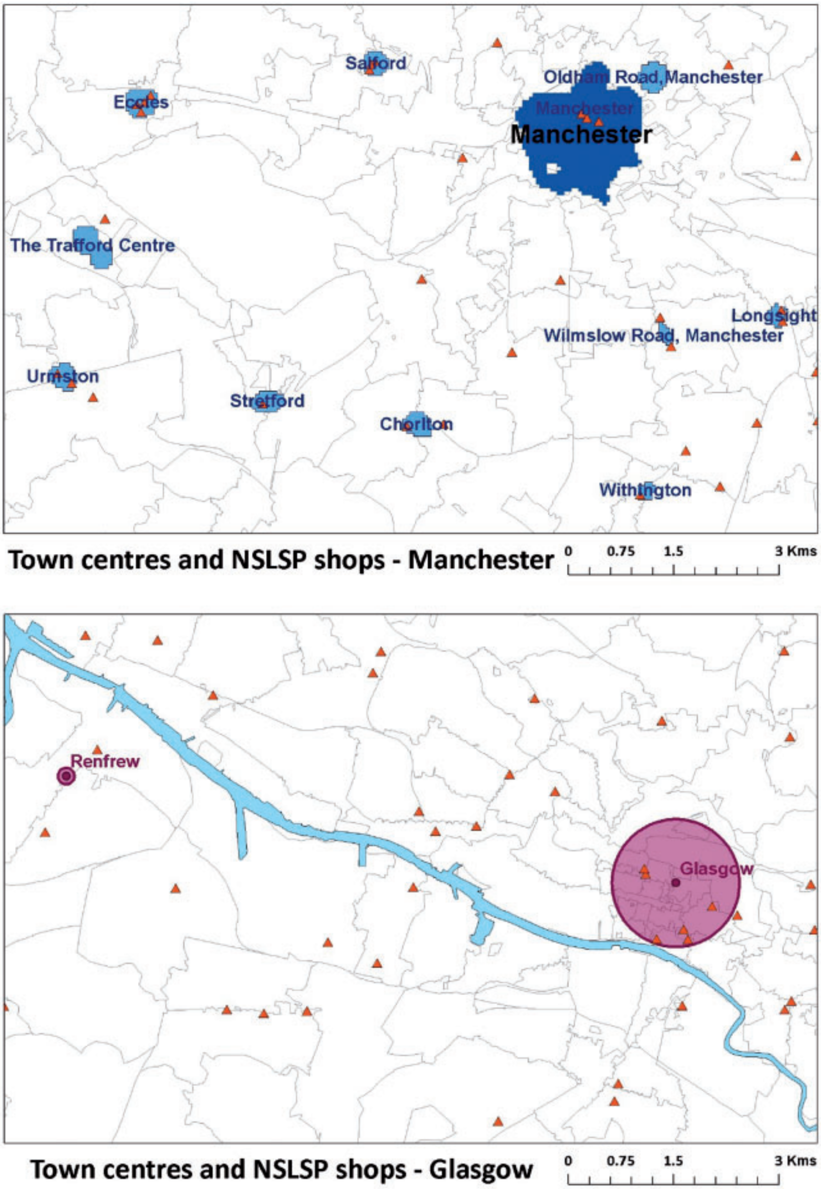


**Figure 2.** Shops inside and outside an ATCA, Manchester, and Glasgow areas.

of the DCLG TCs (which we try to replicate in our methodology), the authors identify four types of factors that characterize TCs: the economy (type and intensity of economic activities), the diversity (of activities carried out in TCs), visitor attractions (transport, retail, and local amenities), and the property of the buildings (intensity and use of land/floorspace of different activities). Unfortunately, we do not have access to the full set of variables used by ODPM/DCLG, so we collected as wide a range as possible—including variables not used in ODPM/DCLG—that we believe capture the four dimensions specified above.

In addition, we use some features related to the physical geography of the TCs and their relative geographical location such as their elevation and distance to the coastline. We also exploit the postcode hierarchy, which in the UK traditionally relates to historical TCs.[17] The different sets of explanatory variables and their main sources are summarized in Table 1.

The variables we use include factors related to the concentration of retail and shopping activity (two data sets from CBRE: the NSLSP and Retail Locations (RETLOC)); size of the retail sector (units and employment); socio-economic and workplace-based factors (including diversity of employees by occupation);[18] infrastructure endowments; local amenities (cultural, consumption, institutional); postcode centrality (based on the order of the postcodes within the postal sector, district and town);[19] location (distance to social and natural amenities); topological features (elevation and slope); and nightlights brightness intensity.[20] We calculated these features around 1–3 km of both the DCLG and OSC TCs, but in this article, we focus on the results using 1 km. To account for non-linearities, some of the variables are included in levels and also with second- and third-order polynomials.

In Step 4 we use all the variables from Table 1 to predict the (log) radius of the DCLG TCs, and after checking how good the fit is (in Step 5), we apply the estimated coefficient to data around the OSC TCs. Formally, our prediction is in two steps. First, we estimate the extent of TCs regressing the explanatory variables (such as shoppers, socio-economic, etc.) on the TC radius using the DCLG England and Wales TCs sample (DCLG TCs):

$$\log(TC\ radius_{DCLG}) = \alpha + \beta_{1,DCLG} shoppers_{DCLG} + \beta_{2,DCLG}\ socioeconomic_{DCLG} + \ldots + \varepsilon. \tag{1}$$

The results of the regressions on the six DCLG 2000 samples explained above are provided in Table 2. Most of the estimates are significantly different from 0 (and by groups, all the sets of explanatory variables are jointly significant) and the goodness of fit of the models is very high ($R^2$ between 0.78 and 0.88). This suggests that our models predict the extent of TCs relatively well.

Having estimated these models, we can save the resulting coefficient values and apply them to different values of the explanatory variables. We do that in Steps 5 and 6. In Step 5, we apply

17 See for example https://www.bph-postcodes.co.uk/guidetopc.cgi

18 Diversity of activities in TCs is one of the key factors determining town centredness in the ODPM methodology. As much as we would like, the data we can use to take into account the diversity of employment around the TC locations (based on the Census 2001 workplace statistics) do not have detailed information on sectors, just occupations. We try to capture diversity in the regressions in two ways: by constructing a normalised Herfindahl (HH) index using the nine occupation categories (which estimates one coefficient) or by flexibly including the share of each occupation on total employment as separate variables, either all nine categories or grouped in a few categories, which allows us to estimate one coefficient per group. In the final results presented in Table 2, we use the second option, and all the coefficients are highly significant. Adding an HH index does not significantly increase the explanatory power of the models.

19 Postal sectors, districts, and towns are aggregations of postcode units in the UK. Their letters and numbers relate to their 'centrality'. For more information, see https://en.wikipedia.org/wiki/Postcodes_in_the_United_Kingdom.

20 We experimented adding additional topological features related to land use (EEA Corine data) and other natural boundaries (share of land in water bodies and green spaces) but none add any further explanatory power to the models.

the coefficients to the DCLG sample to compare the predicted and actual TC radius (and area) for the estimating samples. In Step 6, we apply the coefficients out-of-sample to the set of OSC TCs to predict the extent of TC space for the new set of TC locations. Formally, we calculate the prediction by multiplying the estimated $\hat{\beta}$ s to a different set of locations (OSC TCs):

$$\log(TC\ \widehat{radius}_{OSC}) = \hat{\alpha} + \hat{\beta}_{1,DCLG}\ shoppers_{OSC} + \hat{\beta}_{2,DCLG} socioeconomic_{OSC} + \ldots. \quad (2)$$

Table 3 (and Table A3 for England and Wales separately) summarizes the actual and predicted values for the radius and area of the DCLG TCs for the whole sample (1001, 1075 TCs minus 74 TCs without shops within 1 km of the centroid) for each of the six specifications of Table 2 and for the average of the six predictions. In the bottom panel, for each model, we show the summary statistics of the predictions when we restrict the observations to the sample used in each of the estimated models. By comparing the numbers in each row with the actual values in the first row, we can see that on average, the actual values are very similar to the actual TC values. In Table 3 we provide correlations between the actual and predicted values for the same samples for both England and Wales and separately by country (Table A3). The correlations are again very high, and, in some cases (especially for the predicted area), they are almost equal to 1.

Once we obtain the coefficient in Step 4 and validate the model in Step 5, in Step 6 we apply them to the data around the OSC centroids and calculate their predicted radius of the OSC TCs. This allows us to create buffers around the OSC to draw the extent of the OSC TCs in a map. Figures 3–5 illustrate the method. The DCLG TCs are depicted as solid, irregularly shaped areas, and the OSC TCs centroids are depicted as dark points. The background geographical boundaries correspond to the postal sectors.

Figures 3 and 4 show the steps of the prediction method in three boxes, one for Manchester (in England) and one for Cardiff (in Wales). Box A shows the solid, irregularly shaped DCLG's 'main' 2000 TCs around Manchester (Figure 3) and Cardiff (Figure 4). Then in Box B, the dark points show the OSC TCs centroids of towns and cities around Manchester and Cardiff. Finally, in Box C, our predictions of the extent of TCs around these centroids are seen as shaded dark-bordered circles surrounding Manchester and Cardiff. As explained, these predictions have been obtained by applying the estimated coefficients from Table 2 on the data around the OSC centroids. We can see that for these two cases, the location and extent of both DCLG and OSC TCs are very similar.

Figure 5 shows the predictions around Edinburgh and Glasgow, where there is no DCLG counterpart, since these cities are located in Scotland. As expected, the size of the circles of the two major Scottish cities is larger than those in the neighbouring smaller towns.

There could be several not mutually exclusive reasons accounting for differences between the TCs produced by DCLG and our OSC-predicted samples: (i) the number or location of what are considered towns or cities might differ, (ii) we do not have a comparison group for Scotland, so we cannot check how well the model is doing there; and (iii) the shape of TCs differs (the OSC TCs are circular by construction, while the DCLG TCs can have different shapes, e.g. following a street). However as already discussed and shown in Tables 3 and 4, when we compare the actual and predicted values of the area and radius of the TCs to the DCLG sample, they are actually very similar.

Even if the estimated values of the $R^2$s and the in-sample validations make us confident that we can successfully predict the radius within a TC centroid, we could still be getting

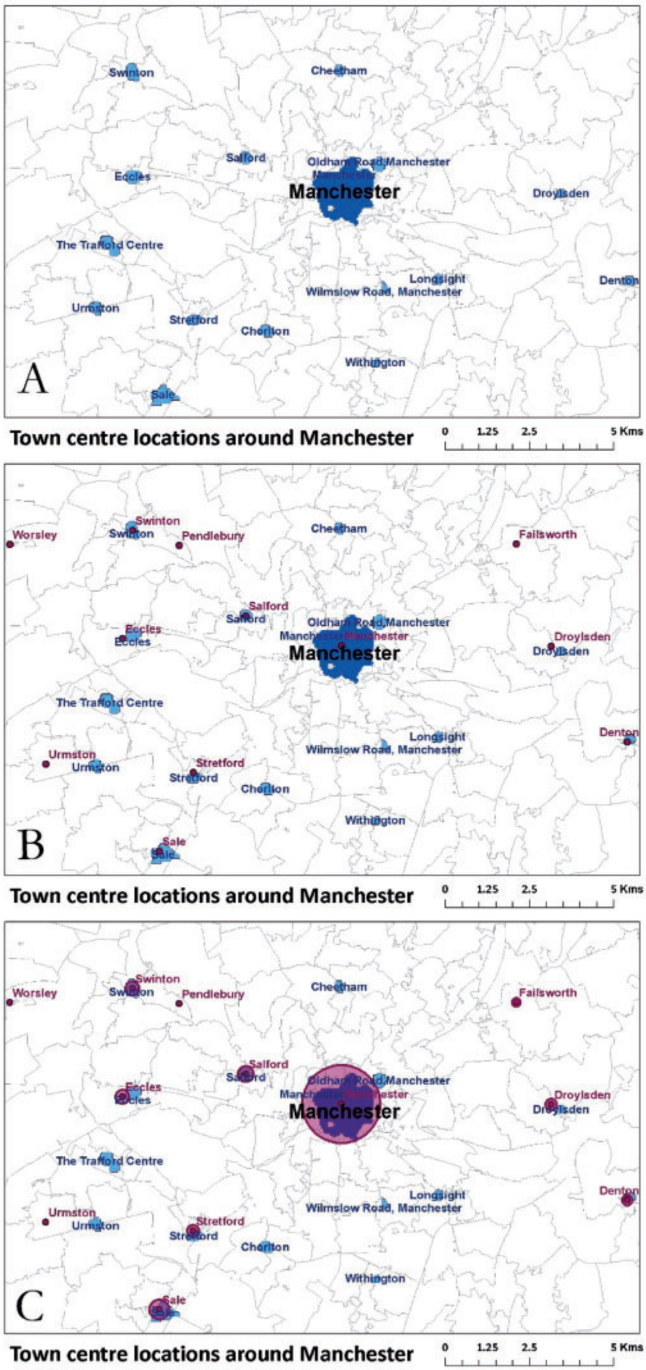


**Figure 3.** TCs prediction, Manchester, England (step-by-step).

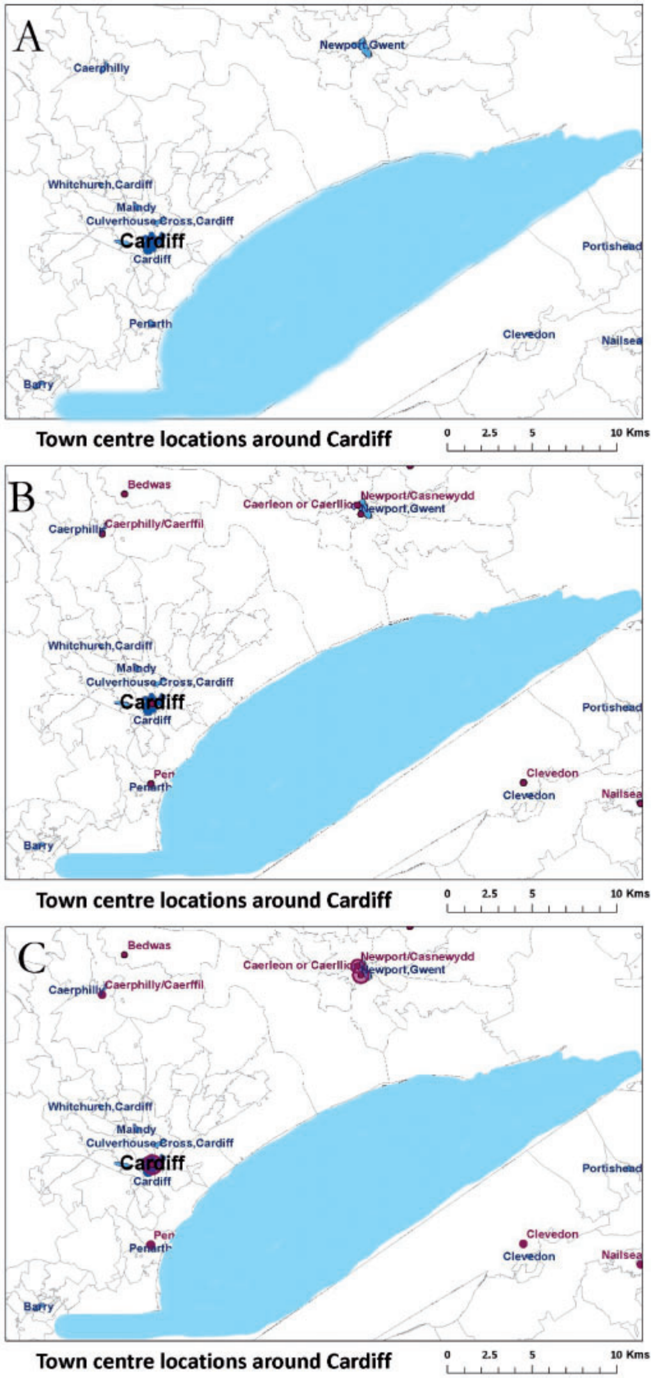


**Figure 4.** TCs prediction, Cardiff, Wales (step-by-step).

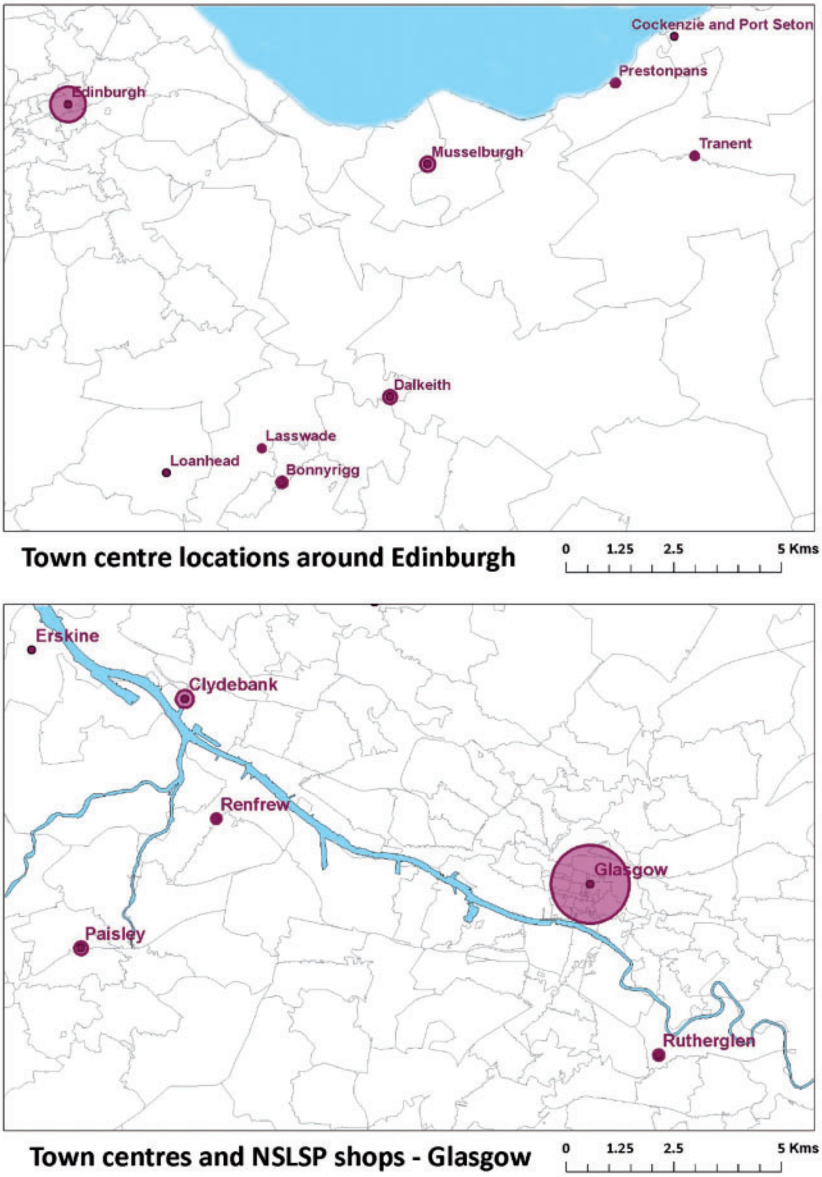


**Figure 5.** TCs prediction, Edinburgh, and Glasgow.

the 'location' of the TCs wrong if the OSC centroids are not sited in the same place as actual TCs. For this reason, in Step 7 we provide a final validation exercise: we compare the socio-economic characteristics of the TCs in the actual DCLG and OSC-predicted samples, first for the countries where we have information for both (England and Wales) and then, for completeness, for Scotland and the whole of Britain. The results of this exercise are shown in Table 5. The table shows the average value for a set of socio-economic and shopping variables using both the DCLG and the OSC samples (we use the criterion of one shop within 500 m of the boundary to select our TCs). These values were obtained combining data from Table 1 and information of the location and extent of the TCs (the original DCLG 2000 shapefiles and the buffered OSC TCs using the average prediction for the six models of Table 2).

The average value of the variable is provided both for its level and for the by-square-kilometre values (to normalize by the size of the TCs and make them more comparable). The DCLG TCs seem to be slightly larger than the OSC ones, especially in Wales, but in general both samples are quite similar. The number of TCs also differs, with more TCs in England in the DCLG sample and fewer in Wales. The last columns show the values for the sample for the whole of Great Britain and for Scotland alone. The Scottish values seem to be somewhere in between the English and the Welsh ones, but they do not look extremely different from the average British or English and Welsh values. In a nutshell, the statistics in Table 5 suggest that the socio-economic and shopping density values of the DCLG and our OSC samples are quite comparable and so we can be reasonably confident that our methodology yields estimates of TCs for all three countries of Great Britain very similar to those of DCLG for England and Wales alone. This opens the door to rigorous analysis of the evolution of TCs in Scotland compared to those in England and Wales and so to an evaluation of policies introduced on one country but not other(s).

## 7. Conclusions

A TC is in a sense the opposite of a pole of inaccessibility. But it is more than that. A TC is a spatial pattern, so it is a recognizable regularity of the urban landscape. Given this one would think it should be central to the research interests of economic geographers. But this interest has not been apparent. In this article, we argue strongly the case for opening the black box of town centredness. Micro-geographical data are now readily available and should be used. In this article we propose and apply a method to exploit this type of data to define the location and extent of TCs in Britain.

This article starts with an apparently naïve question: How can one identify a TC in a given city? The answer proves not to be so simple. To answer it we find we need a whole new method. TC policies have been around for several decades, in many European countries (apart from Britain we mainly discuss the case of The Netherlands). These policies seem to have been applied with less rigour than rhetoric. We cite the case of a handbook for Scottish TCs in which there is no definition at all of what a ‘Town Centre’ is, where it is to be found, or how it is to be defined. There are pictures but no maps or definitions. Our research tries to bridge this gap, proposing a new methodology to locate, identify, delimit, and determine the radius of TCs. Calibrating our model on TCs defined by ODPM (2004), we test our method in a full Great Britain setting, but it is easily transferred to other locations or countries because of its reproducibility and ease of calculation.

In this article, we apply a method for predicting the location and extent of TC space to all of Great Britain. Our method relies on four assumptions. The first assumption is that the DCLG TC definitions are good approximations of the true TCs for England and Wales. The second assumption is that the underlying socio-economic and geographical factors within a radius of around 1 km of the TC centroids are effective determinants of TC space. The validity of this assumption can be assessed by looking at the goodness-of-fit statistics of our models predicting the extent of the DCLG’s TCs and at the evidence provided in Table 3–5. The third assumption is that the OS list of towns and cities provides a reliable set of potential TC locations. The final assumption is that the determinants of TC space in Scotland do not systematically differ from those in England and Wales, both in observed and unobserved characteristics relevant to defining TCs. If all these assumptions hold, we can satisfactorily apply the coefficients on socio-economic and geographical variables estimated in Table 2 to Britain-wide data to yield estimates of the location and extent of TCs for England, Wales, and, in particular, Scotland. Equally, so long as the critical assumptions hold, the methodology could be adapted to identify TCs in other countries.

While this study gives an answer to the question of the extent of TCs and so allows one to estimate where their centroids are located, there is no such thing as *the* answer. As our robustness checks and data validations suggest, the method can be considered 'successful' with a correlation of actual to predicted radius of 0.75–0.99 depending on the sample. Our predictions for England and Wales match the actual DCLG ATCA 2000 quite accurately. In Scotland, its direct accuracy cannot be judged because there are no 'official' DCLG TCs—to offset for which is one of the purposes of this study. However, the exploration of socio-economic and shopping density values in and very close to the TCs defined with both methodologies suggests that they provide a very similar picture. Overall, we judge that our method is promising and certainly provides a useful tool to be applied for the evaluation of TCFP, and more generally, for the evaluation of any policy that applies to TCs.

Our final aim is policy discussion and evaluation. Having workable and agreed definitions of TCs and their boundaries is a necessary step if we are to have an open, consistent, and reliable discussion or evaluation of relevant policy. TCs as a distinct spatial pattern of modern cities deserve this effort. In this article, we hope we are demonstrating a replicable method for the analysis of this particular spatial organization which will help in policy development and analysis. At least, the discussion both in the UK and in Europe signals an urgent need to first consider town centredness seriously as a precondition to policy analysis and debate.

TCs, their extent and the hierarchy of TCs, are, as we argue in Section 2, closely related to, indeed an extension of, CPT and gravity models. Many of our assumptions are borrowed directly from these two intellectual traditions, but some come from a more empirical approach where several of the recent papers we discuss have shown the way. We hope future policy debates may incorporate our primary aim: that we should have agreed definitions of things before we launch discussion, let alone policy for them; and perhaps borrow or adapt our methodology. TCs should be recognized as real entities with real shapes, with real areas and real boundaries, capable of real descriptions and definitions.

Finally, with this article we aim to contribute to improving not just debates about cities and TCs but to other debates where people and policy proceed ahead of any clear definition of what it is they are analysing or generating policy for. This has very much been the case with TCs (as it has with other concepts relating to urban development such as 'sprawl'), but they are not abstruse ideas, and we hope this article has shown that they can be clearly and unambiguously defined and identified and that they are basically material, applied, and experimental in nature.

## Acknowledgements

The authors thank SERC/CEP for funding (ESRC grant ES/M010341/1), and R.S.-G. acknowledges support from the British Academy. The authors thank Mr Mark Teale from CBRE and DCLG for providing them with data. ABI data accessed via NOMIS under BRES notice NTC/BRES15-P0572. The authors benefited from comments and suggestions from participants at the SERC Work-in-progress research seminar, at the CESifo Conference on the Use of Geo-coded Data in Economic

Research and at the University of Birmingham City-REDI seminar. The authors finally thank Steve Gibbons and Daniel Arribas-Bel for comments and an anonymous referee for helpful criticisms. Any errors remain our own.

## References

Cheshire, P. C., C. A. L. Hilber, and I. Kaplanis (2015), "Land Use Regulation and Productivity-Land Matters: Evidence from a Uk Supermarket Chain", *Journal of Economic Geography*, **15**, 43–73.

Cheshire, P. C., C. A. L. Hilber, and R. Sanchis-Guarner (2014), "Do English Planning Policies Make Shopping More Sustainable?", Working paper, IEB Urban Economics Workshop, Barcelona, May 2014. http://www.ieb.ub.edu/files/PapersWSUE2014/Sanchis.pdf (last accessed 26 February 2018).

Christaller, W. (1933), *Die Zentralen Orte in Süddeutschland*. G. Fischer, Jena.

Denike, K. G. and J. B. Parr (1970), "Production in Space, Spatial Competition, and Restricted Entry", *Journal of Regional Science*, **10**, 49–64.

Department of the Environment (1996), *Planning Policy Guidance: Town Centres and Retail Developments. Revised PPG6*. HMSO, London.

Dicken, P. and P. E. Lloyd (1990), *Location in Space: Theoretical Perspectives in Economic Geography*, 3rd edn. Harper and Row, New York.

DETR (1998), *Town Centres: Defining Boundaries for Statistical Monitoring - Feasibility Study*. HMSO, London.

Dolega, L., M. Pavlis, and A. Singleton (2016), "Estimating Attractiveness, Hierarchy and Catchment Area Extents for a National Set of Retail Centre Agglomerations", *Journal of Retailing and Consumer Services*, **28**, 78–90.

Everitt, B. and T. Hothorn (2011), *An Introduction to Applied Multivariate Analysis with R*. Springer, New York.

Evers, D. (2002), "Window on The Netherlands. The Rise (and Fall?) of National Retail Planning", *Tijdschrift Voor Economische En Sociale Geografie*, **93**, 107–13.

Fujita, M., P. R. Krugman, and A. Venables (1999), *The Spatial Economy: Cities, Regions and International Trade*. MIT Press, Cambridge, Massachusetts.

Guimarães, P. (2016), "Revisiting Retail Planning Policies in Countries of Restraint of Western Europe", *International Journal of Urban Sciences*, **20**, 361–80.

Huff, D. L. (1963), "A Probabilistic Analysis of Shopping Center Trade Areas", *Land Economics*, **39**, 81–90.

Huff, D. L. (1964), "Defining and Estimating a Trading Area", *Journal of Marketing*, **28**, 34–8.

Huff, D. L. (2003), "Parameter Estimation in the Huff Model", *ArcUser*, 34–6.

Klaesson, J. and Ö. Öner (2014), "Market Reach for Retail Services", *The Review of Regional Studies*, **44**, 153–76.

Lösch, A. (1940), *Die Räumliche Ordnung Der Wirtschaft*. G. Fischer, Jena.

ODPM (2004), *Producing Boundaries and Statistics for Town Centres, England and Wales 2000 Interim Report*. Office of the Deputy Prime Minister, London.

ODPM and CASA (2002), *Producing Boundaries and Statistics for Town Centres - London Pilot Study Summary Report*. The Stationery Office, London.

Pavlis, M., L. Dolega, and A. Singleton (2017), "A Modified DBSCAN Clustering Method to Estimate Retail Center Extent", *Geographical Analysis*, doi: 10.1111/gean.12138.

Reilly, W. J. (1929), *Methods for the Study of Retail Relationships*. University of Texas, Austin.

Reilly, W. J. (1931), *The Law of Retail Gravitation*. Knickerbocker Press, New York.

Thurstain-Goodwin, M. and D. Unwin (2000), "Defining and Delineating the Central Areas of Towns for Statistical Monitoring Using Continuous Surface Representations", *Centre for Advanced Spatial Analysis*, **18**, 1–11.

## Appendix

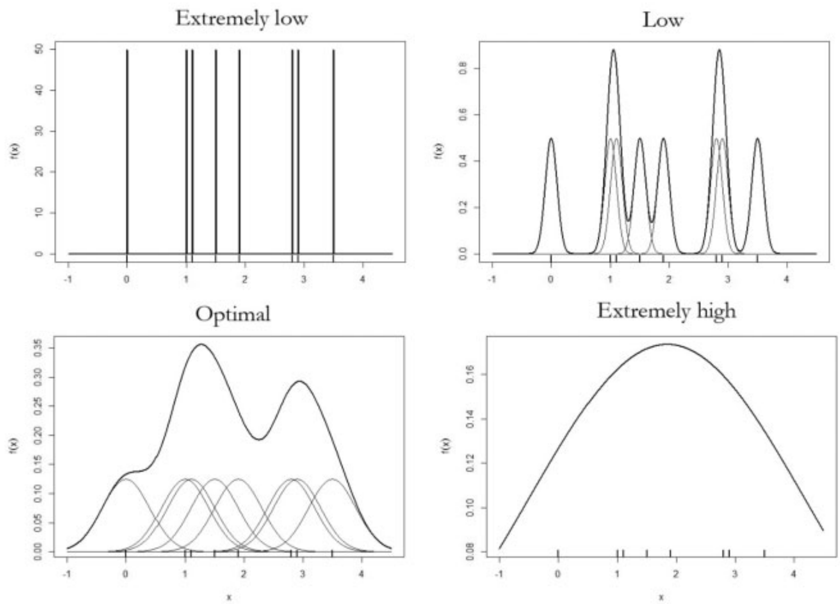


**Figure A1.** Bandwidth selection (modified from Everitt and Hothorn 2011).

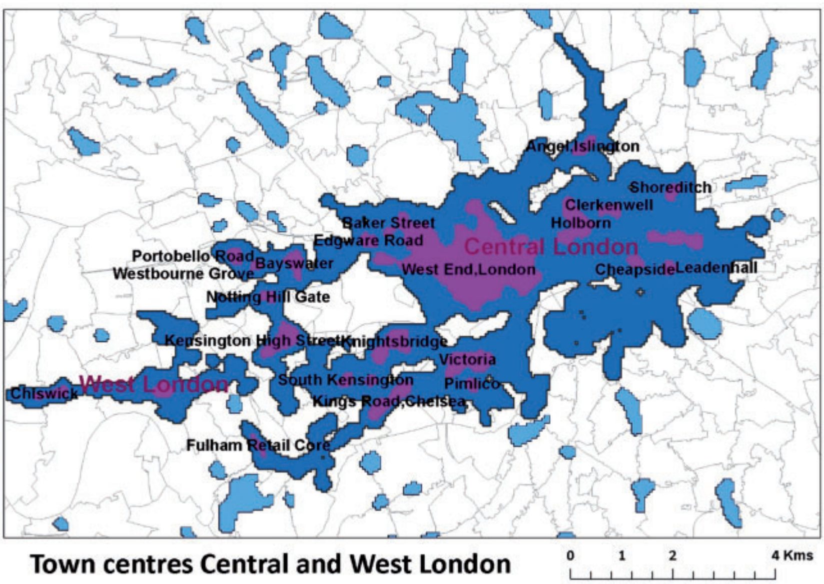


**Figure A2.** ATCA and RCs in London in the DCLG sample.

**Table A1.** List of abbreviations

| | |
|---|---|
| ABI | Annual Business Inquiry |
| ATCAs | Areas of Town Centre Activity |
| BRES | Business Register and Employment Survey |
| CASA | Centre for Advanced Spatial Analysis |
| CBRE | CBRE Group., Inc. |
| CEP | Centre for Economic Performance |
| City-REDI | City-Region Economic and Development Institute |
| CPT | Central Place Theory |
| DCLG | Department for Communities and Local Government |
| DETR | Department of the Environment, Transport and the Regions |
| EEA | European Environment Agency |
| ESRC | Economic and Social Research Council |
| GB | Great Britain |
| GIS | Geographic Information System |
| NASA | National Aeronautics and Space Administration |
| NGDC | National Geophysical Data Center |
| NOAA | National Oceanic and Atmospheric Administration |
| NOMIS | Official Labour Statistics Portal (from ONS) |
| NSLSP | National Survey on Local Shopping Patterns |
| NSPD | National Statistics Postcode Directory |
| ODPM | Office of the Deputy Prime Minister |
| ONS | Office for National Statistics |
| OS | Ordnance Survey |
| OSC | Towns and cities list in the Ordnance Survey Gazetteer |
| POI | Points of Interest |
| RCs | Retail Cores |
| RETLOC | Retail Locations |
| SERC | Spatial Economics Research Centre |
| TC(s) | Town Centre(s) |
| TCFP | Town Centre First Policies |
| UPC | Unit Post Code |
| UK | United Kingdom |
| VOA | Valuation Office Agency |

**Table A2.** Details on variables used in the estimates of Table 2 (all within 1 km of centroid)

| Variable | Description | Data source |
|---|---|---|
| *Number of (grocery) shoppers* | Shoppers in shops located around the centroid from shopping patterns survey | NSLSP and RETLOC 1998 |
| *Number of (grocery) shops* | Number of grocery shops located around the centroid from both CBRE surveys/directories | NSLSP and RETLOC 1998 |
| *Average distance to (grocery) shops* | Average distance from centroid to all shops around it from both CBRE surveys/directories | NSLSP and RETLOC 1998 |
| *Distance to closest (grocery) shop* | Distance from centroid to closest shop | NSLSP 1998 |
| *Number of units in retail sector* | Establishments in sector 52 (SIC 1992) | NOMIS ABI 1998 |
| *Employment in retail sector* | Employment in sector 52 (SIC 1992) | NOMIS ABI 1998 |
| *Density of total population* | Total residential population in a given area | UK Census 2001 |
| *Workplace employment* | Workplace-based total employment | UK Census 2001 |
| *Share of workplace employment in different occupation levels (diversity)* | High (managerial/professional), medium (intermediate/supervisory), low (routine) | UK Census 2001 |
| *Average commuting distance* | Average distance in kilometres | UK Census 2001 |
| *% of commuters by foot or bike* | Percentage using bike or walking to work | UK Census 2001 |
| *% of commuters using public transport* | Percentage using bus/tube/rail/taxi | UK Census 2001 |
| *Average household size* | Average number of members in household | UK Census 2001 |
| *Average age of the resident population* | Average age of all resident population | UK Census 2001 |
| *% of population aged 18–44 years* | Percentage of residents aged 18–44 years | UK Census 2001 |
| *Total unemployment rate* | Total unemployed as percentage of active population | UK Census 2001 |
| *Share of students in population* | Share of students in population | UK Census 2001 |
| *Share of retirees in population* | Share of retired workers in population | UK Census 2001 |
| *Kilometres of all-type roads* | Motorways, primary, A & B, and minor roads | OS Strategi 2009 |
| *Number of bus stations* | Number of major bus stations (not bus stops) | OS Points of Interest 2015 |
| *Number of tube/tram stations* | Number of tube and tram stations | OS Points of Interest 2015 |
| *Number of rail stations* | Number of rail station (not tube or tram) | OS Points of Interest 2015 |

(continued)

**Table A2.** Continued

| Variable | Description | Data source |
|---|---|---|
| *Number of libraries* | Number of public libraries | OS Points of Interest 2015 |
| *Number of museums* | Number of museums open to the public | OS Strategi 2009 |
| *Number of art galleries* | Number of art galleries | OS Points of Interest 2015 |
| *Number of cinemas and theatres* | Number of cinemas and theatres | OS Points of Interest 2015 |
| *Number of discos and nightclubs* | Number of discos and nightclubs | OS Points of Interest 2015 |
| *Number of landmarks* | Landmark point (historic building/castles, etc.) | OS Strategi 2009 |
| *Number of cafes, restaurants, and pubs* | Number of cafes, restaurants, and pubs | OS Points of Interest 2015 |
| *Number of B&B, hotels, and motels* | Private accommodation establishments | OS Points of Interest 2015 |
| *Number of youth hostels* | Youth hostel establishments | OS Points of Interest 2015 |
| *Number of local government sites* | Number of government-related buildings | OS Points of Interest 2015 |
| *Number of tourist information offices* | All tourist offices (including seasonal) | OS Strategi 2009 |
| *Number of visitor centres* | All visitor centres which are not tourist offices | OS Strategi 2009 |
| *Share of central addresses (postal sector in postal town)* | Central addresses (those whose postal sector number is 1) in postal town: total number | NSPD and Wikipedia |
| *Distance to first postcode in postal town* | First postcode in the postcode hierarchy | NSPD and Wikipedia |
| *Distance to closest town hall* | Town hall belonging to district administration | Wikipedia |
| *Distance to closest rail or tube station* | Closest tube or rail (one or the other) | OS Points of Interest 2015 |
| *Distance to closest point in coastline* | Closest distance to the coastline | OS Strategi 2009 |
| *Distance to closest river or lake* | Closest distance to water bodies | OS Strategi 2009 |
| *Distance to closest natural park or woodland* | Closest distance to natural national parks or woodland (green spaces) | OS Strategi 2009 |
| *Elevation/terrain slope statistics* | Standard deviation, mean, max, range of terrain elevation (altitude), and slope (degrees) | OS Panorama 50x50 m |
| *Nightlight brightness statistics* | Standard deviation, mean, max, range, and sum of nightlight brightness (non-censored version) | NASA and NOAA-NGDC |

**Table A3.** England and Wales TC radiuses, actual and predicted

| | TC derived radius in km | | | | | TC area in $km^2$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Observed | Mean | Standard deviation | Minimum | Maximum | Observed | Mean | Standard deviation | Minimum | Maximum |
| England | | | | | | | | | | |
| Observed value | 944 | 0.248 | 0.156 | 0.113 | 3.187 | 944 | 0.270 | 1.077 | 0.040 | 31.905 |
| Average all samples | 944 | 0.263 | 0.168 | 0.107 | 2.973 | 944 | 0.316 | 1.021 | 0.036 | 28.018 |
| Sample 1 betas | 944 | 0.276 | 0.200 | 0.104 | 3.095 | 944 | 0.366 | 1.257 | 0.034 | 30.091 |
| Sample 2 betas | 944 | 0.274 | 0.199 | 0.106 | 3.076 | 944 | 0.360 | 1.237 | 0.035 | 29.718 |
| Sample 3 betas | 944 | 0.263 | 0.181 | 0.105 | 3.051 | 944 | 0.320 | 1.089 | 0.034 | 29.245 |
| Sample 4 betas | 944 | 0.259 | 0.150 | 0.105 | 2.526 | 944 | 0.282 | 0.738 | 0.034 | 20.040 |
| Sample 5 betas | 944 | 0.257 | 0.155 | 0.107 | 2.711 | 944 | 0.283 | 0.831 | 0.036 | 23.088 |
| Sample 6 betas | 944 | 0.252 | 0.166 | 0.106 | 3.382 | 944 | 0.285 | 1.218 | 0.035 | 35.929 |
| Sub-sample 1 | 768 | 0.258 | 0.160 | 0.104 | 3.095 | 768 | 0.289 | 1.120 | 0.034 | 30.091 |
| Sub-sample 2 | 820 | 0.252 | 0.156 | 0.106 | 3.076 | 820 | 0.276 | 1.075 | 0.035 | 29.718 |
| Sub-sample 3 | 876 | 0.246 | 0.150 | 0.111 | 3.051 | 876 | 0.260 | 1.025 | 0.039 | 29.245 |
| Sub-sample 4 | 780 | 0.253 | 0.121 | 0.105 | 1.040 | 780 | 0.247 | 0.311 | 0.034 | 3.397 |
| Sub-sample 5 | 832 | 0.248 | 0.119 | 0.107 | 1.070 | 832 | 0.237 | 0.308 | 0.036 | 3.600 |
| Sub-sample 6 | 894 | 0.241 | 0.113 | 0.108 | 1.030 | 894 | 0.222 | 0.287 | 0.036 | 3.336 |

(continued)

**Table A3.** Continued

| Variable | TC derived radius in km | | | | | TC area in $km^2$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Observed | Mean | Standard deviation | Minimum | Maximum | Observed | Mean | Standard deviation | Minimum | Maximum |
| Wales | | | | | | | | | | |
| Observed value | 57 | 0.218 | 0.096 | 0.113 | 0.613 | 57 | 0.178 | 0.198 | 0.040 | 1.180 |
| Average all samples | 57 | 0.216 | 0.095 | 0.128 | 0.570 | 57 | 0.175 | 0.199 | 0.052 | 1.022 |
| Sample 1 betas | 57 | 0.219 | 0.097 | 0.125 | 0.585 | 57 | 0.180 | 0.206 | 0.049 | 1.077 |
| Sample 2 betas | 57 | 0.217 | 0.098 | 0.128 | 0.590 | 57 | 0.178 | 0.209 | 0.051 | 1.095 |
| Sample 3 betas | 57 | 0.213 | 0.091 | 0.122 | 0.534 | 57 | 0.168 | 0.182 | 0.047 | 0.895 |
| Sample 4 betas | 57 | 0.219 | 0.098 | 0.128 | 0.588 | 57 | 0.181 | 0.209 | 0.052 | 1.087 |
| Sample 5 betas | 57 | 0.217 | 0.098 | 0.127 | 0.589 | 57 | 0.178 | 0.210 | 0.050 | 1.089 |
| Sample 6 betas | 57 | 0.213 | 0.090 | 0.123 | 0.532 | 57 | 0.167 | 0.179 | 0.048 | 0.888 |
| Sub-sample 1 | 44 | 0.218 | 0.093 | 0.125 | 0.538 | 44 | 0.218 | 0.093 | 0.125 | 0.538 |
| Sub-sample 2 | 50 | 0.212 | 0.089 | 0.128 | 0.538 | 50 | 0.212 | 0.089 | 0.128 | 0.538 |
| Sub-sample 3 | 55 | 0.208 | 0.082 | 0.122 | 0.495 | 55 | 0.208 | 0.082 | 0.122 | 0.495 |
| Sub-sample 4 | 44 | 0.218 | 0.093 | 0.128 | 0.538 | 44 | 0.218 | 0.093 | 0.128 | 0.538 |
| Sub-sample 5 | 50 | 0.212 | 0.089 | 0.127 | 0.535 | 50 | 0.212 | 0.089 | 0.127 | 0.535 |
| Sub-sample 6 | 55 | 0.207 | 0.081 | 0.123 | 0.494 | 55 | 0.207 | 0.081 | 0.123 | 0.494 |

*Notes*: Samples define as explained in notes of Table 3.

*CESifo Economic Studies*, 2018, 327–338
doi: 10.1093/cesifo/ifx026
Advance Access Publication Date: 8 January 2018
Original article

# GIS for Credible Identification Strategies in Economics Research

**Masayuki Kudamatsu***

*Osaka School of International Public Policy (OSIPP), Osaka University, Toyonaka, Osaka, Japan. e-mail: m.kudamatsu@gmail.com

## Abstract

This article surveys the use of geographic information systems (GIS) for the credible identification of causal impacts in recent economics research. It describes how each geo-processing tool in GIS allows economists to use data on geography and weather as sources of exogenous variation for estimating the impact of various 'treatments'. The diverse range of treatments discussed in this survey includes disease, school competition, land suitability for agriculture, infrastructure, the elasticity of housing supply, mass media, learning from friends, slave trade, the appropriability of crop harvests, and terrain ruggedness. (JEL codes: C88, O10)

**Key words**: econometric and statistical methods, economic development

## 1. Introduction: Why Geographic Information System for Economists?

In the past decade or so, the use of geographic information systems (GIS) has increasingly become popular among economists. There are at least two reasons for this popularity. First, GIS allows economists to observe the previously unobserved. Satellite images, which can be handled with GIS, allow us to observe what slips from official statistics due to illegality (e.g. deforestation in Indonesia; Burgess et al. 2012) or due to the lack of state capacity (e.g. gross domestic product (GDP) growth, which can be approximated by changes in the intensity of night-time light; Henderson et al. 2012).[1] Scanned old maps, which can be converted into statistics with GIS, also allow us to observe historical data such as the building of roads since the year of independence in Kenya (Burgess et al. 2015) and the location of ethnic groups in Africa during the period of the Atlantic slave trade (Nunn 2008). As a result, previously infeasible pieces of empirical research can now be conducted.

The second reason for the popularity of GIS is that it helps economists identify causal impacts in a more credible way than previously possible. Data sets on geography and

1 See Donaldson and Storeygard (2016) for an overview of how satellite images have been used in economics research.

weather can be handled easily with GIS, and they are great sources of exogenous variation in various treatments of interest. This survey focuses on this aspect of the use of GIS in economics research.

Below I organize the survey by geo-processing tools in GIS. Each geo-processing tool takes spatial data sets as inputs and produces a new spatial data set or a table of statistics. While GIS software provides myriads of them, some geo-processing tools are more relevant for economists than others. I hope this survey will be helpful for those wishing to use GIS for economics research to prioritize what to learn.

The next section discusses geo-processing tools to merge different data sets by location. Section 3 introduces tools to process elevation data, followed by Section 4 where the tools involving distance calculation are covered. Section 5 describes the tools to conduct cell-by-cell calculations from raster data sets. Section 6 concludes.

## 2. Merging Data sets

Almost every piece of empirical research involves merging different data sets. Three major geo-processing tools for this purpose are Spatial Join, Intersect, and Zonal Statistics.

### 2.1 Spatial Join

The Spatial Join tool allows us to merge two data sets by location, rather than by the name or identifier of observations. The tool is particularly attractive for economists because it allows standard socio-economic data sets (e.g. household and firm surveys, administrative data at the level of sub-national districts) to be merged with weather data.

Since weather is beyond the control of human beings, it provides exogenous regressors or instrumental variables to identify the causal impact of weather itself (see Dell et al. 2014 for a survey) or weather-induced events such as disease prevalence (Kudamatsu et al. 2016) and income shock in agricultural society (Miguel et al. 2004). Many data sets on weather, however, adopt a point location on the Earth's grid system as the unit of observation. Without GIS, it is painstaking to match each socio-economic observation (households, firm plants, sub-national districts, countries) with its nearest grid point in the weather data set.

Perhaps one of the best applications of the Spatial Join tool in economics research is by Alsan (2015), who empirically assesses whether the abundance of tsetse flies has hindered historical and contemporary economic development in Africa. Found only in Africa, tsetse flies kill domesticated animals, which were in human history an important input to agricultural production as a source of draft power and manure for fertilizer. Thus, some historians and biologists argue that tsetse flies were historically the reason for Africa's low agricultural productivity and, as a result, low population density.

Proving this causal impact of tsetse flies is empirically challenging. We need data on the abundance of tsetse flies across different parts of Africa. Such historical data are hard to come by, and even if it is available, the abundance of tsetse flies is endogenous to socio-economic conditions: more affluent societies may have managed to contain the prevalence of tsetse flies.

To overcome this empirical challenge, Alsan (2015) exploits the biological fact that the birth rate and mortality of tsetse flies depend on temperature and humidity. She constructs an index of the climatic suitability for tsetse flies' survival from data on humidity and temperature in the year of 1871, the earliest year for which meteorologists 'forecast' daily weather from the climate model at each of the two-degree cells across Africa (Compo et al. 2011). She then uses the Spatial Join tool to merge the index with the geo-coded anthropological data on ethnic groups across pre-colonial Africa (Murdock 1967). Regression analysis shows that the more suitable the climate is for tsetse flies, the less likely the ethnic group used domesticated animals and the lower the population density in the area inhabited by the ethnic group.

Regarding the question of whether and how tsetse flies have an adverse effect on contemporary economic development in Africa, Alsan (2015) also finds that ethnic groups in an area more suitable for tsetse flies had a lower level of political centralization. Since a higher degree of the pre-colonial political centralization is known to predict a higher level of economic development in Africa (Gennaioli and Rainer 2007, Michalopoulos and Papaioannou 2013), Alsan's (2015) findings suggest that the historical abundance of tsetse flies is one source of underdevelopment in today's Africa.

### 2.2 Intersect

The Intersect tool merges several spatial data sets (polygons and polylines) by location. Unlike the Spatial Join tool, however, the tool creates the intersections of input polygons and polylines as its outputs. For example, the output from merging river polylines with district polygons is a set of river segments divided by the district boundaries. From this output, it is easy to count the number of rivers within each district, either using the Dissolve tool or reading the output data table in statistical software.

An application of this tool in economics is Hoxby (2000). She asks whether competition improves the quality of schools in the USA. Competition is measured by the number of schools in the same metropolitan area. However, an area may have more schools because, for example, the local government allocates more of its budget to education, which itself can directly affect the quality of schools. If so, the correlation between the number of schools and their quality does not reflect the causal impact of competition.

Hoxby (2000) uses the number of streams within each metropolitan area as an instrumental variable for the number of schools, arguing that the boundary of school districts was historically drawn along streams. She does find a positive correlation between the number of streams and the number of schools, and the instrumental variable estimation shows that more competition leads to a higher quality of schools (measured by test scores, school graduates' income at the age of 32 years, etc.).

The same identification strategy is used by Bai and Jia (2016). They ask whether the abolishment of the 1300-year-old civil exam system in Imperial China in 1905 caused uprisings that led to the Revolution of 1911. Each prefecture of China was assigned a fixed number of candidates who could pass the exam, mostly based on the number of counties within the prefecture. Since Chinese county boundaries often coincided with streams, the number of streams within a prefecture can be used as an instrument for how many people could enter the civil service each year. The instrumental variable estimation shows that the higher the quota the prefecture had, the more uprisings took place. Additional evidence suggests that the main reason for this causal link is the complaint of those would-be political elites who were now denied upward mobility through the civil exam.

### 2.3 Zonal Statistics

The Zonal Statistics tool merges polygons with raster data by taking the summary statistics of raster cell values for each polygon. The tool is usually used to aggregate gridded data, often derived from satellite images (such as elevation and night-time light), into the existing administrative units around the world (countries, provinces, counties, and so forth).

Michalopoulos (2012) takes full advantage of the Zonal Statistics tool by conducting what he calls cross-'virtual country' regression. He asks whether the pre-historic cause of ethnic diversity is the variability in agricultural suitability of land. He hypothesizes that an ethnic group was formed by a group of people specializing in the same method of production, the latter of which was determined by the degree of land suitability for agriculture. This hypothesis was important for economists to test because ethnic diversity had been treated as exogenous (i.e. not affected by socio-economic factors) and believed to be the cause of underdevelopment (Easterly and Levine 1997), a low quality of government (Alesina et al. 2003), and civil wars (Montalvo and Reynal-Querol 2005).

Ethnic diversity and land quality variability are indeed positively correlated at the country level. However, the boundary of countries is likely to be endogenous to ethnic diversity. For example, an ethnically homogeneous area with homogenous land quality may have formed a country first, and the remaining areas may have become a country later, with heterogeneity both in ethnicity and in land quality as a result.

To deal with this concern, Michalopoulos (2012) uses the Zonal Statistics tool to measure the standard deviation of land quality within each of 'virtual countries', that is, 2.5-degree cells across the world. The number of languages spoken in each virtual country is obtained by the Intersect tool to merge the virtual country polygons with the map of language groups across the world (World Language Mapping System 2006). Since the boundaries of virtual countries are arbitrarily drawn, whatever heterogeneity within each of them does not depend on the human history of state formation. As each country can contain many virtual countries, country fixed effects can also be controlled for.

He finds a positive correlation between variability in land quality and the number of languages spoken in each virtual country, conditional on country fixed effects. Thus, we can interpret this correlation as the causal impact of variability in land quality on ethnic diversity. Without the Zonal Statistics tool, this empirical methodology would not have been feasible.

## 3. Elevation

Elevation data[2] have been one of the most popular spatial data sets for economists because socio-economic factors in most cases do not alter a location's altitude. While GIS has several tools for processing elevation data, the most popular among economists have been the Slope tool and the Irregular Terrain Model (ITM).

2 The 30 arc-second resolution version of NASA's Shuttle Radar Topography Mission (SRTM30) has been the most widely used elevation data among economists. The data are downloadable at dds.cr.usgs.gov/srtm/version2_1/SRTM30/. See Farr et al. (2007) for detail.

### 3.1 Slope

The Slope tool takes elevation raster data as an input and calculates the slope of the terrain for each cell in the input raster data. The output can be expressed as the percent rise or as the degree of gradient.

It is one of the most common geo-processing tools used by economists. Dinkelman (2011) and Qian (2008) use the district average land slope as an instrument for electrification in South Africa (flatter land is cheaper to lay power lines) and for tea production in China (hilly terrain is suitable for tea cultivation), respectively, to estimate the impact on female labour participation (electrification relieves women from household chore) and on the male-to-female ratio (tea production makes women more valuable because it is difficult for men to pick tea leaves without damaging it).

Saiz (2010) uses the Slope tool to identify the land unsuitable for residential development due to a slope above 15%. He then calculates the percentage of undevelopable areas in each US city and shows that the elasticity of housing prices with respect to housing demand shocks is higher for a city with a higher percentage of undevelopable areas. This study arms economists with an instrumental variable for the supply of houses: Diamond (2016), Harari (2016), and Chen and Kung (2016) all use the percentage of land with a slope above 15% as an instrument for housing rent across cities in the USA, city shapes in India, and land sales revenue by local governments in China, respectively.

Duflo and Pande (2007) ask whether the construction of dams reduced poverty in India over the period of 1971–1999. Since the location of new dams can politically be chosen to target poorer areas, they instrument the number of dams in a district with the fractions of river segments whose gradient is 1.5–3, 3–6, and over 6%. The rationale behind this first-stage relationship is that, from an engineering point of view, the construction of dams is easier with moderately sloped rivers (gradient 1.5–3%) for irrigation dams and with very steep rivers (gradient over 6%) for hydroelectricity dams. Their instrumental variable estimation reveals that the construction of dams reduced poverty for districts downstream from dams by increasing agricultural production due to irrigation. But it also increased poverty for districts where dams were built because of an increased volatility of agricultural production caused by the salinization of soil and the restricted use of water.

### 3.2 Irregular Terrain Model

A more advanced use of elevation data for economics research is to use the ITM to predict the signal reception of television or radio channels (Hufford 2002). The ITM takes, as inputs, the location and height of signal transmitting antennas and the elevation data to calculate the spatial distribution of signal loss.

The signal strength across different locations can then be used as an instrument to identify the impact of the mass media. Using this identification strategy, Olken (2009) shows that an increased number of television channels reduced social capital (measured by participation in group activities) in Indonesia. Yanagizawa-Drott (2014) finds that one cause of the Rwandan genocide in 1994 is the Anti-Hutu propaganda broadcast by a radio channel.

As exemplified by these pieces of research, the recent rise of the empirical literature on the impact of mass media (see Prat and Strömberg 2013 for a survey) is partly due to the availability of data on geographically determined (thus exogenous) signal strength thanks to GIS.

## 4. Distance

The calculation of distance per se does not necessarily require the use of GIS. The distance between two point locations can be obtained from the great-circle distance formula.[3] The power of GIS comes from its ability to identify all the other observations within a certain distance from each observation (the Buffer tool) and to calculate the shortest distance from a point location to the nearest polyline or polygon (the Near tool).[4]

### 4.1 Buffer

The Buffer tool creates a circular polygon of a chosen radius around each of the input point locations. Using the Spatial Join tool to merge these circular 'buffer' polygons with the original point locations, researchers can easily identify each observation's neighbours within a radius of any length.

This tool advances the methodology of peer effect estimation. As Manski (1993) shows, the question of whether the behaviour of someone is affected by that of his or her friends poses a significant empirical challenge because friends typically share the same environment and the same information, which itself may bring about similar behaviour among friends.

Conley and Udry (2010) overcome this empirical challenge thanks to GIS. Interested in whether farmers in rural Ghana learn from their friends about the use of fertilizer for cultivating pineapples (a new crop introduced in the surveyed villages), they measure the location of each farmer's plot with a GPS receiver[5] and identify each plot's geographic neighbours with the Buffer tool. They then control for the average use of fertilizer in the neighbouring plots within a radius of 1 km, to take into account the common productivity shock among neighbours. They also measure whom each farmer talks to about farming, some of whom cultivate plots more than 1 km away. This way, they differentiate changes in behaviour caused by their friend's behaviour from those in response to common productivity shocks shared by all geographic neighbours. They demonstrate that the failure to control for common productivity shocks would lead to the spuriously estimated presence of peer effect for labour use in the cultivation of cassava, the traditional crop that everyone in the village knows how to grow.

### 4.2 Near

The Near tool matches each input point location in one data set with its nearest observation in another data set and also calculates the distance to it. If observations in the second data set are polylines or polygons, the tool also identifies the nearest point on the line or on the polygon boundary.

3 Stata has a module called globdist to implement this calculation.

4 A number of studies in economics have used distance itself as a source of exogenous variation in treatments. See Gibson and McKenzie (2007) for a discussion of these studies and of the validity of such identification strategy.

5 Burgert et al. (2013) introduce how GPS receivers should be used in surveys.

Nunn (2008) asks whether slave trade has a long-lasting impact on economic development in Africa today. Since slave trade may be endogenous to the historical level of development, he instruments the number of exported slaves from each country with the distance to the nearest slave trade centre in the Americas. The Near tool helps construct this instrument. It first finds the nearest point on the coast of the African continent from each country's centroid (and the distance between them). The tool then finds the nearest slave trade centre to each of these nearest points on the coast (and the distance between them). Using the sum of these distance measures as an instrument, Nunn (2008) finds that countries with more slaves exported are poorer today.[6]

The Near tool is also useful for spatial regression discontinuity design (Keele and Titiunik 2015, Keele et al. 2017), by calculating the distance from each observation to the treatment area boundary (with negative values indicating the outside of the treatment area). Researchers can then check if the outcome, when plotted against the distance to the treatment area boundary, discontinuously changes at 0. As long as it is plausibly exogenous which side of the boundary the unit of observation is located on, a discontinuous change of the outcome at the boundary can be interpreted as a causal impact of the treatment.

The reduction of a two-dimensional vector of each location (i.e. longitude and latitude) into the scalar distance to the boundary, however, means that a treated observation close to one segment of the border will be compared to a control observation close to a different segment of the border. These two observations cannot be seen as similar because their locations are not very close to each other. To avoid such comparisons, the nearest boundary segment indicator, easily obtained by the Near tool, is used to categorize observations that are close to each other.[7]

As a pioneering study using the spatial regression discontinuity design, Dell (2010) finds that the forced labour system during the Spanish colonial rule, implemented in specific areas of the Andes for mining silver and mercury, had a negative impact on today's living standards in Peru. Another example is Michalopoulos and Papaioannou (2014), who show that a difference in the quality of national governments (such as the rule of law and corruption) in Africa does not translate into changes in living standards across national borders.

## 5. Map Algebra

Map Algebra conducts cell-by-cell calculation across several raster data sets. Each raster data acts like a variable in algebra. Suppose that two raster data files are given the names of 'ras1' and 'ras2'. Then typing ras1 + 2*ras2 + 3 in Python code yields a new raster data set

6 In a follow-up piece of research, Nunn and Wantchekon (2011) use the distance to the coastline from each country's centroid as an instrument to estimate the impact of slave trade on African people's trust in others and in governments today.

7 Dividing the treatment area boundary into several segment polylines is tricky with ArcGIS. One approach is to use the 'positionAlongLine' method to obtain the coordinate of a point that divides a boundary polyline into halves, thirds, quarters, etc. Then the Split Line at Point tool can be used to create boundary segment polylines.

whose cell value is the sum of the 'ras1' cell value, the 'ras2' cell value doubled, and three. Researchers can also calculate summary statistics across raster data sets for each location by using the Cell Statistics tool or across neighbouring cells within the same raster data set by using the Focal Statistics tool.

### 5.1 Cell Statistics

A recent study by Mayshar et al. (2015) uses Map Algebra and the Cell Statistics tool at their best. They hypothesize that the cultivation of cereal crops gave rise to the emergence of states while growing tubers and root crops discouraged it. Storable and easily appropriable foods were necessary for a state to emerge because the collection of foods as taxation would otherwise be too costly. Cereal crops are storable, and their appropriation is not costly, as they are visible above the ground and their harvest season is a particular period of the year. Tubers and root crops are, on the other hand, perishable, grown underground, and harvested throughout the year.

To provide evidence for their hypothesis on state formation, Mayshar et al. (2015) use the Global Agro-Ecological Zones (GAEZ) data set (Fischer et al. 2012) on potential yields for various crops by rain-fed agriculture. This data set is available as a set of 5 arc-minute resolution rasters, constructed from the crop growth model with climate and soil data as inputs. Since the potential yields in this data set are predicted from factors beyond human control (biology, climate, and terrain), economists regard them as exogenous determinants of what crops are cultivated across the world.[8]

Using Map Algebra, Mayshar et al. (2015) first multiply the potential yield with the calorie conversion factor for each crop. Then using the Cell Statistics tool, they obtain the maximum potential caloric yields by cereal crops and by tubers and root crops in each location. With Map Algebra, they then take the difference between the maximum cereal caloric yield and the maximum root crop caloric yield as a measure of the suitability for state formation.

To identify the causal impact, they exploit the Columbian Exchange: around the year of 1500, some cereal and root crops from the New World became available in the Old World (maize, cassava, white potato, and sweet potato), while others from the Old World became available in the New World (various cereal crops and yams). Thus, the difference in the maximum caloric yield between cereal and root crops exogenously changed around the year of 1500. Mayshar et al. (2015) then find that an increase in the relative caloric productivity of cereal crops against root crops does predict the emergence of states.[9]

8 Several recent studies in economics use the GAEZ data for various purposes: to show the impact of potato cultivation on population in the Old World (Nunn and Qian 2011); to predict changes in comparative advantage of crop production across countries due to climate change and how agricultural trade patterns will change accordingly (Costinot et al. 2016); and to find the positive impact of genetically modified soybean seeds on industrial growth through the release of agricultural labour in the areas of Brazil where soybeans are more cultivable (Bustos et al. 2016).

9 The same identification strategy is also used by Galor and Özak (2016). They show that an increased caloric productivity thanks to the Columbian Exchange made residents more patient because, with a higher return to agricultural production, patience is more rewarding.

### 5.2 Focal statistics

In the context of quantifying topographic heterogeneity in wild habitats, Riley et al. (1999) proposed the terrain ruggedness index. It is defined as the square root of the sum of squared differences in altitude from all the eight neighbouring locations:

$$TRI_{xy} = \sqrt{\sum_{i=x-d}^{x+d} \sum_{j=y-d}^{y+d} (e_{ij} - e_{xy})^2},$$

where $x$ and $y$ indicate the longitude and latitude of the location for which the index is constructed, $d$ the spatial resolution of elevation data (e.g. 30 arc-second for SRTM30), and $e_{ij}$ the altitude of the location with longitude $i$ and latitude $j$. The Focal Statistics, together with Map Algebra, can compute this index. To see this, first observe:

$$\begin{aligned} TRI_{xy} &= \sqrt{\sum_i \sum_j (e_{ij} - e_{xy})^2} \\ &= \sqrt{\sum_i \sum_j e_{ij}^2 - 2e_{xy} \sum_i \sum_j e_{ij} + 9e_{xy}^2}. \end{aligned}$$

Map Algebra computes $e_{ij}^2$. Then the Focal Statistics tool calculates $\sum_i \sum_j e_{ij}^2$ and $\sum_i \sum_j e_{ij}$. The remaining subtraction, addition, and multiplication can be done with Map Algebra.

Nunn and Puga (2012) calculate the ruggedness index for each 30 arc-second cell across the world and use its country-level average as an exogenous regressor in a cross-country regression of GDP per capita today. They show that ruggedness reduces GDP except in Africa, where ruggedness is associated with higher GDP. They argue that, while ruggedness increases the cost of transportation and consequently hinders economic development, it played a different role in Africa, where rugged terrain made slave raids difficult during the period of slave trade. Since slave trade negatively affects Africa's development (as discussed in Section 4.2 above), ruggedness in Africa has saved residents from the long-run negative consequence of slave trade.

## 6. Conclusion

This survey overviews how GIS has been applied to obtain more credible estimates of causal impacts in recent economics research. For many economists, an initial investment to learn how to use GIS software may appear prohibitively costly. However, mastering the use of GIS is mostly about knowing which geo-processing tool(s) to use for a given purpose. This survey showcases the mapping from each tool to the purpose of research.[10]

10 The remaining thing to learn is to write Python code to implement geo-processing tools for the ease of replication. For that matter, I have developed self-contained tutorials, based on the PhD course that I gave for the past several years. They are accessible online at sites.google.com/site/mkudamatsu/gis. One of the students who took this course (Rogall 2014) has written a paper using the Near, Buffer, and Spatial Join tools to reveal that another factor of the 1994 Rwandan genocide was the presence of armed groups who encouraged and forced civilians to participate in mass killing.

The geo-processing tools mentioned in this survey are relatively simple ones, quickly becoming the standard tools for empirical economists. More advanced tools are beginning to be embraced by economists. Examples include the shortest path algorithm of Dijkstra (1959) (Donaldson 2012, Faber 2014, Dell 2015, Storeygard 2016, Yamasaki 2016), spatial statistics (Harari 2016), and minimum bounding geometry (König et al. 2017). These and other advanced geo-processing tools will be likely to spur innovative research in economics in the coming years.

## Acknowledgements

This article was prepared for the CESifo Economic Studies Conference on the Use of Geo-coded Data in Economic Research, Munich, 18 and 19 November 2016. This article originates from several years of teaching how to use ArcGIS for applied econometric research to PhD students at Stockholm University and Stockholm School of Economics (2010–2013), University of Copenhagen (2013), and University of Gothenburg (2016). The author is grateful to Matz Dahlberg for providing an opportunity to write this survey, John Östh for invaluable comments from the viewpoint of a geographer, and an anonymous referee for helpful comments and suggestions.

## References

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat and R. Wacziarg (2003), "Fractionalization", *Journal of Economic Growth* **8**, 155–194.

Alsan, M. (2015), "The Effect of the TseTse Fly on African Development", *American Economic Review* **105**, 382–410.

Bai, Y. and R. Jia (2016), "Elite Recruitment and Political Stability: The Impact of the Abolition of China's Civil Service Exam", *Econometrica* **84**, 677–733.

Burgess, R., M. Hansen, B. A. Olken, P. Potapov and S. Sieber (2012), "The Political Economy of Deforestation in the Tropics", *Quarterly Journal of Economics* **127**, 1707–1754.

Burgess, R., R. Jedwab, E. Miguel, A. Morjaria and G. Padró I Miquel (2015), "The Value of Democracy: Evidence from Road Building in Kenya", *American Economic Review* **105**, 1817–1851.

Burgert, C. R., B. Zachary and J. Colston (2013), "Incorporating Geographic Information into Demographic and Health Surveys: A Field Guide to GPS Data Collection", in *Demographic and Health Surveys Manual 9*. dhsprogram.com/publications/index.cfm

Bustos, P., B. Caprettini and J. Ponticelli (2016), "Agricultural Productivity and Structural Transformation: Evidence from Brazil", *American Economic Review* **106**, 1320–1365.

Chen, T. and J. K.-S. Kung (2016), "Do Land Revenue Windfalls Create a Political Resource Curse? Evidence from China", *Journal of Development Economics* **123**, 86–106.

Compo, G. P., J. S. Whitaker, P. D. Sardeshmukh, N. Matsui, R. J. Allan, X. Yin, B. E. Gleason, R. S. Vose, G. Rutledge, P. Bessemoulin, S. Brönnimann, M. Brunet, R. I. Crouthamel, A. N. Grant, P. Y. Groisman, P. D. Jones, M. C. Kruk, A. C. Kruger, G. J. Marshall, M. Maugeri, H. Y. Mok, Ø. Nordli, T. F. Ross, R. M. Trigo, X. L. Wang, S. D. Woodruff and S. J. Worley (2011), "The Twentieth Century Reanalysis Project", *Quarterly Journal of the Royal Meteorological Society* **137**, 1–28.

Conley, T. and C. Udry (2010), "Learning about a New Technology: Pineapple in Ghana", *American Economic Review* **100**, 35–69.

Costinot, A., D. Donaldson and C. Smith (2016), "Evolving Comparative Advantage and the Impact of Climate Change in Agricultural Markets: Evidence from 1.7 Million Fields around the World", *Journal of Political Economy* **124**, 205–248.

Dell, M. (2010), "The Persistent Effects of Peru's Mining Mita", *Econometrica* **78**, 1863–1903.

Dell, M. (2015), "Trafficking Networks and the Mexican Drug War", *American Economic Review* **105**, 1738–1779.

Dell, M., B. F. Jones and B. A. Olken (2014), "What Do We Learn from the Weather? the New Climate-Economy Literature", *Journal of Economic Literature* **52**, 740–798.

Diamond, R. (2016), "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000", *American Economic Review* **106**, 479–524.

Dijkstra, E. (1959), "A Note on Two Problems in Connection with Graphs", *Numerische Mathematik* **1**, 269–271.

Dinkelman, T. (2011), "The Effects of Rural Electrification on Employment: New Evidence from South Africa", *American Economic Review* **101**, 3078–3108.

Donaldson, D. (2012), "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure", *American Economic Review*, forthcoming.

Donaldson, D. and A. Storeygard (2016), "The View from Above: Applications of Satellite Data in Economics", *Journal of Economic Perspectives* **30**, 171–198.

Duflo, E. and R. Pande (2007), "Dams", *Quarterly Journal of Economics* **122**, 601–646.

Easterly, W. and R. Levine (1997), "Africa's Growth Tragedy: Policies and Ethnic Divisions", *Quarterly Journal of Economics* **112**, 1203–1250.

Faber, B. (2014), "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System", *Review of Economic Studies* **81**, 1046–1070.

Farr, T. G., P. A. Rosen, E. Caro, R. Crippen, R. Duren, S. Hensley, M. Kobrick, M. Paller, E. Rodriguez, L. Roth, D. Seal, S. Shaffer, J. Shimada, J. Umland, M. Werner, M. Oskin, D. Burbank and D. Alsdorf (2007), "The Shuttle Radar Topography Mission", *Reviews of Geophysics* **45**, RG2004.

Fischer, G., F. O. Nachtergaele, S. Prieler, E. Teixeira, G. Toth, H. v. Velthuizen, L. Verelst and D. Wiberg (2012), Global Agro-Ecological Zones (GAEZ v3.0): Model Documentation, gaez.fao.org

Galor, O. and Ö. Özak (2016), "The Agricultural Origins of Time Preference", *American Economic Review* **106**, 3064–3103.

Gennaioli, N. and I. Rainer (2007), "The Modern Impact of Precolonial Centralization in Africa", *Journal of Economic Growth* **12**, 185–234.

Gibson, J. and D. McKenzie (2007), "Using Global Positioning Systems in Household Surveys for Better Economics and Better Policy", *World Bank Research Observer* **22**, 217–241.

Harari, M. (2016), Cities in Bad Shape: Urban Geometry in India, real.wharton.upenn.edu/~harari/research.html

Henderson, J. V., A. Storeygard and D. N. Weil (2012), "Measuring Economic Growth from Outer Space", *American Economic Review* **102**, 994–1028.

Hoxby, C. M. (2000), "Does Competition among Public Schools Benefit Students and Taxpayers?", *American Economic Review* **90**, 1209–1238.

Hufford, G. (2002), The ITS Irregular Terrain Model, Version 1.2.2: The Algorithm, www.its.bldrdoc.gov/media/50676/itm_alg.pdf

Keele, L. and R. Titiunik (2015), "Geographic Boundaries as Regression Discontinuities", *Political Analysis* **23**, 127–155.

Keele, L., S. Lorch, M. Passarella, D. Small and R. Titiunik (2017), "An Overview of Geographically Discontinuous Treatment Assignments with an Application to Children's Health Insurance", *Advances in Econometrics* **38**, 147–194.

König, M. D., D. Rohner, M. Thoenig and F. Zilibotti (2017), "Networks in Conflict: Theory and Evidence from the Great War of Africa", *Econometrica* **85**, 1093–1132.

Kudamatsu, M., T. Persson and D. Strömberg (2016), Weather and Infant Mortality in Africa, sites.google.com/site/mkudamatsu/

Manski, C. F (1993), "Identification of Endogenous Social Effects: The Reflection Problem", *Review of Economic Studies* **60**, 531–542.

Mayshar, J.,O. Moav, Z. Neeman and L. Pascali (2015), "Cereals, Appropriability and Hierarchy", CEPR Discussion Paper, no. 10742, Centre for Economic Policy Research, London, UK.

Michalopoulos, S (2012), "The Origins of Ethnolinguistic Diversity", *American Economic Review* **102**, 1508–1539.

Michalopoulos, S. and E. Papaioannou (2013), "Pre-Colonial Ethnic Institutions and Contemporary African Development", *Econometrica* **81**, 113–152.

Michalopoulos, S. and E. Papaioannou (2014), "National Institutions and Subnational Development in Africa", *Quarterly Journal of Economics* **129**, 151–213.

Miguel, E., S. Satyanath and E. Sergenti (2004), "Economic Shocks and Civil Conflict: An Instrumental Variables Approach", *Journal of Political Economy* **112**, 725–753.

Montalvo, J. G. and M. Reynal-Querol (2005), "Ethnic Polarization, Potential Conflict, and Civil Wars", *American Economic Review* **95**, 796–816.

Murdock, G. P. (1967), *Ethnographic Atlas*. Pittsburgh, PA: University of Pittsburgh Press.

Nunn, N. (2008), "The Long-Term Effects of Africa's Slave Trades", *Quarterly Journal of Economics* **123**, 139–176.

Nunn, N. and D. Puga (2012), "Ruggedness: The Blessing of Bad Geography in Africa", *Review of Economics and Statistics* **94**, 20–36.

Nunn, N. and N. Qian (2011), "The Potato's Contribution to Population and Urbanization: Evidence from a Historical Experiment", *Quarterly Journal of Economics* **126**, 593–650.

Nunn, N. and L. Wantchekon (2011), "The Slave Trade and the Origins of Mistrust in Africa", *American Economic Review* **101**, 3221–3252.

Olken, B. A. (2009), "Do Television and Radio Destroy Social Capital? Evidence from Indonesian Villages", *American Economic Journal: Applied Economics* **1**, 1–33.

Prat, A. and D. Strömberg (2013), "The Political Economy of Mass Media", in Advances in Economics and Econometrics: Theory and Applications, Proceedings of the Tenth World Congress of the Econometric Society. Cambridge University Press, Cambridge, UK.

Qian, N. (2008), "Missing Women and the Price of Tea in China: The Effect of Sex-Specific Earnings on Sex Imbalance", *Quarterly Journal of Economics* **123**, 1251–1285.

Riley, S. J., S. D. DeGloria and R. Elliot (1999), "A Terrain Ruggedness Index That Quantifies Topographic Heterogeneity", *Intermountain Journal of Sciences* **5**, 23–27.

Rogall, T. (2014), Mobilizing the Masses for Genocide, sites.google.com/site/thorstenrogall

Saiz, A. (2010), "The Geographic Determinants of Hosing Supply", *Quarterly Journal of Economics* **125**, 1253–1296.

Storeygard, A. (2016), "Farther on down the Road: Transport Costs, Trade and Urban Growth in Sub-Saharan Africa", *Review of Economic Studies* **83**, 1263–1295.

World Language Mapping System. (2006), World Language Mapping System, Version 3.2. www.gmi.org/wlms

Yamasaki, J. (2016), Railroads, Technology Adoption, and Modern Economic Development: Evidence from New Data in the Late 19th - Early 20th Centuries Japan, www.junichiyamasaki.com

Yanagizawa-Drott, D. (2014), "Propaganda and Conflict: Evidence from the Rwandan Genocide", *Quarterly Journal of Economics* **129**, 1947–1994.

# CESifo Economic Studies

## SUBSCRIPTIONS continued

**For further information, please contact:** Journals Customer Service Department, Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, UK. Email: jnls.cust.serv@oup.com. Tel (and answerphone outside normal working hours): +44 (0)1865 353907. Fax: + 44 (0)1865 353485. **In the US, please contact:** Journals Customer Service Department, Oxford University Press, 2001 Evans Road, Cary, NC 27513, USA. Email: jnlorders@oup.com. Tel (and answerphone outside normal working hours): 800 852 7323 (toll-free in USA/Canada). Fax: 919 677 1714. **In Japan, please contact:** Journals Customer Service, Oxford University Press, 4-5-10-8F Shiba, Minato-ku, Tokyo 108-8386, Japan. Tel. +81 3 5444 5858. Fax. +81 3 3454 2929. E-mail: custserv.jp@oup.com.

**Methods of payment**. (i) Cheque (payable to Oxford University Press, to Oxford University Press, Cashiers Office, Great Clarendon Street, Oxford OX2 6DP, UK) in GB£ Sterling (drawn on a UK bank), US$ Dollars (drawn on a US bank), or EU€ Euros. (ii) Bank transfer to Barclays Bank Plc, Oxford Group Office, Oxford (bank sort code 20-65-18) (UK), overseas only Swift code BARC GB 22 (GB£ Sterling to account no. 70299332, IBAN GB89BARC20651870299332; US$ Dollars to account no. 66014600, IBAN GB27BARC20651866014600; EU€ Euros to account no. 78923655, IBAN GB16BARC20651878923655). (iii) Credit card (Mastercard, Visa, Switch or American Express).

## PERMISSIONS



## ADVERTISING

Inquiries about advertising should be sent to CESifo Economic Studies, Oxford Journals Advertising, Journals Division, Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, UK. Email: jnlsadvertising@oup.com. Tel: +44 (0)1865 353329, (0)1865 354637 or (0)1865 354767. Fax: +44 (0)1865 353774.

## DISCLAIMER



## INSTRUCTIONS TO AUTHORS

*CESifo Economic Studies* is an economics journal circulated to a wide audience including governments, businesses, and policy-makers as well as academics and students. Authors should make every possible effort to avoid technicalities in the form of either terminology which will be unfamiliar to a general readership or mathematical presentations. Where specialist terminology is used, it should be defined clearly and examples of its application given. The use of mathematics should be avoided. Authors who do employ mathematics in their articles should ensure that it is a necessary component of their presentation which clarifies an unfamiliar concept. If mathematics is used, authors should restrict it to an appendix as much as possible.

**Format.** On the first manuscript page provide the title, names of all authors, affiliations, any footnotes to the title, address to which proofs are to be sent, a short running title, and both fax number and e-mail address of the author to be contacted.

**Abstracts.** Authors are asked to provide a short abstract in English, which should not exceed 100 words. Please supply up to five keywords, and the Journal of Economic Literature index number.

**Submitting your paper.** Once you have prepared your manuscript according to the instructions above please visit the online submission web site: http://mc.manuscriptcentral.com/cesifo

**Refereeing.** All articles submitted to the journal will be refereed.

For detailed Instructions to Authors, please visit the journal's website: http://academic.oup.com/cesifo



Typeset by Cenveo publisher services, Bangalore, India
Printed by Bell and Bain Ltd, Glasgow, UK

OXFORD
UNIVERSITY PRESS

## About CESifo Economic Studies

CESifo Economic Studies aims to bridge the gap between specialised research in economics and the interest of non-specialists. It publishes policy oriented articles written in a style accessible to all economists regardless of their specialisation