Cipullo, Davide; Colombo, Luca V.A.; Magnani, Michele; Onorato, Massimiliano Gaetano

Working Paper

Historical Newspaper Markets

CESifo Working Paper, No. 12194

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Cipullo, Davide; Colombo, Luca V.A.; Magnani, Michele; Onorato, Massimiliano Gaetano (2025) : Historical Newspaper Markets, CESifo Working Paper, No. 12194, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/333740

CES ifo Working Papers

12194
2025

October 2025

Historical Newspaper Markets

Davide Cipullo, Luca V.A. Colombo, Michele Magnani,
Massimiliano Gaetano Onorato

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Historical newspaper markets*

Davide Cipullo† Luca V. A. Colombo‡ Michele Magnani§

Massimiliano Gaetano Onorato¶

September 2025

Abstract

This paper proposes a novel methodology to identify the geographic market of local newspapers when information on their diffusion is not available or is not sufficiently granular. We illustrate the methodology using historical data from 154 newly digitized newspapers published in Italy between 1919 and 1922. Combining machine learning-augmented optical character recognition techniques, multi-way fixed-effect regressions, and GIS tools, our approach allows us to estimate markets based on news content. Text-based location of newspaper markets considerably improves over assuming that market boundaries coincide with administrative aggregations. We discuss how our technique strengthens the usage of newspapers as a granular and time-varying source of historical information and offers new avenues for identification strategies.

JEL Classification: C18, C81, N01, N94
Keywords: newspaper markets, media coverage, text analysis, inter-war Italy

*We thank participants to seminars held at Università Cattolica del Sacro Cuore, University of Bologna, and ASREC Europe Conference 2025 for their comments and suggestions. All errors are our own. Davide Ciocia, Lucrezia Ferrari, and Paolo Iovino provided excellent research assistance. Massimiliano Onorato is grateful to the Whitney & Betty MacMillan Center for International and Area Studies for support and hospitality while working on this project and acknowledges financial support from European Union Funding - NextGenerationEU -Public Notice No. 104 of February 2, 2022, issued by the Ministry of University and Research for the submission of project proposals under "PRIN 2022", funded within the framework of the National Recovery and Resilience Plan (NRRP) - Mission 4: "Education and Research" - Component 2: "From Research to Business" - Investment 1.1: "Fund for the National Research Programme (NRP) and Projects of Significant National Interest (PRIN)". Funding Approval Decree No. 967 of 30.06.2023 - Project Title: "Empirical Evaluation of Historical Policies: Fascist Land Reclamation Program and Human Capital Accumulation in the Long-Run" - MUR Project Code: 20227SNASS - CUP: J53D23004120006.

†Department of Economics and Finance, Università Cattolica del Sacro Cuore; CESifo; CIFREL; Uppsala Centre for Fiscal Studies. Address: Largo A. Gemelli 1, 20123 Milan (Italy). E-mail: davide.cipullo@unicatt.it.

‡Department of Economics and Finance, Università Cattolica del Sacro Cuore. Address: Largo A. Gemelli 1, 20123 Milan (Italy). E-mail: lucava.colombo@unicatt.it.

§Department of Economics, University of Bologna. Address: Piazza A. Scaravilli 2, 40126 Bologna (Italy). E-mail: michele.magnani9@unibo.it.

¶Department of Economics, University of Bologna. Address: Piazza A. Scaravilli 2, 40126 Bologna (Italy). E-mail: massimiliano.onorato@unibo.it.

1 Introduction

Newspapers are an important source of information across various strands of literature in social sciences, and advances in newspaper digitization are likely to further increase their relevance for research in the near future. Their appeal is evident: they are available at high frequency (typically daily), register events at fine geographic scales, and often provide information where administrative records do not. This is especially relevant from a historical perspective (see, e.g., Beach and Hanlon, 2023b; Ferrara et al., 2024), as newspapers are often the most easily accessible source of data, when not the only one available at a granular and high-frequency level.

Newspapers are often used to measure economic variables in a specific area or across localities (see Calderon et al. (2023), Feigenbaum and Gross (2024), Gentzkow et al. (2011, 2014), and Snyder and Strömberg (2010) for prominent examples). When using newspapers to capture information at the local level, identifying the 'market' of each newspaper is crucial to avoid the risk of misinterpreting the information contained in the news. Readership surveys or official statistics about newspaper sales are rarely available or representative at a sufficiently granular level to assist in this task – nowadays and even more so for historical sources (Drago et al., 2014; Gentzkow et al., 2014; Scheve and Serlin, 2025). In the literature, it is customary to assume that the market of each newspaper coincides with a fixed area around the headquarters of that newspaper: for example, Gentzkow et al. (2014) refers to the headquarters' city; Djourelova et al. (2024) to counties; George and Waldfogel (2006) and Gentzkow and Shapiro (2010) to larger administrative aggregations such as the headquarters' commuting zones or states. This approach is likely to identify the market of a newspaper in an imprecise way only. Out of the several newspapers that are published in the same city at the same time, some may have a very localized interest – their market being a city or even just a neighborhood – while others may have a much broader focus. The former might be a reliable data source about a city, but may not offer adequate information at the state level. The latter may, instead, offer comprehensive information about the state, but not necessarily about the city in which they are headquartered. In both cases, identifying a newspaper market with exogenously assigned geographic boundaries provides a biased representation of the socio-economic forces that are relevant for a specific territory.

This paper develops a novel text-based methodology, which can be easily applied to different historical contexts and countries, to identify newspaper markets when circulation data are unavailable or too coarse. Our approach leverages the supply of news content to estimate newspaper markets allowing for shapes of different sizes that depart from administrative boundaries. This flexibility is valuable for three reasons. First, obtaining endogenously varying market sizes allows us to extract reliable and appropriately geo-referenced information from historical newspapers, and to select those among them

focusing on a specific territory. Second, measuring newspaper markets with an approach that does not force them to be nested in administrative aggregations allows us to design credible identification strategies for causal analysis. For example, our methodology enables comparisons within-newspaper and time (across administrative borders) that would be otherwise unfeasible.[1] Third, our method allows us to assess the competitive structure of local media more precisely.

The first step of our methodology consists of retrieving the raw frequency of mentions of each locality in the textual content of newspapers, obtained by digitizing their scanned images by means of *LayoutParser* (Shen et al., 2021). We then regress raw frequencies on city-time, newspaper-time, and city-newspaper fixed effects. Such regression models allow us to dispose of all newspaper-time specific shocks stemming from a newspaper production technology (e.g., number of pages, frequency, word counts, and editorial standpoints), as well as of the city-time shocks stemming from the overall coverage of a locality or event, common across newspapers.[2] Finally, we refine the sets of city-newspaper tuples identified from the above estimation (i.e., the set of cities on which a specific newspaper focuses) using GIS tools to finalize the reconstruction of historical newspaper markets as clusters of contiguous localities.

Although our methodology is entirely general and applicable to various countries and time periods, we illustrate it by focusing on the Italian press between 1919 and 1922. This is an especially challenging period for our purposes, as it is marked by rapid increases in literacy, the proliferation and decline of numerous newspapers, and intense political instability eventually culminating in the rise of fascism. The markets resulting from the application of our method to a set of 154 daily or weekly newspapers (including newspapers of linguistic minorities) display several desirable properties. We confirm that broad-based and locally focused newspapers coexist, and markets do not necessarily overlap with administrative units. In fact, comparing our estimated markets with those proxied by administrative aggregations, we find that only a small fraction of markets can be properly approximated using at least one level of aggregation of administrative units. Moreover, some markets are better identified by large administrative aggregations (e.g., regions, comparable to U.S. commuting zones), while others by more granular units.

[1] A prominent exception using contemporary data is Snyder and Strömberg (2010), who exploit the fact that U.S. congressional districts are not nested within counties (their proxy for newspaper markets). Similar analyses are unfeasible if electoral districts overlap with or are nested in administrative aggregations.

[2] The raw frequency of mentions bears the risk of confounding national salience with local attention and that of obtaining text measures that are contingent on outlet-specific production technologies. Many localities are likely to be mentioned in the text of a local newspaper just because some events of national interest occurred there or because the newspaper contains a large amount of textual context and is published very frequently. In contrast, some localities within the area of interest of a newspaper may not be mentioned just because there were no significant events or because of capacity constraints. Puglisi and Snyder (2011) and Calderon et al. (2023) provide examples of studies that leverage the reporting of events that occur outside a newspaper market.

To evaluate the practical improvements in measuring economic variables when using our approach, we compare a proxy for the illiteracy rate at the municipal level based on the textual content of newspapers with the actual illiteracy rates reported in the 1921 Census.[3] We find that the proxy obtained by using our methodology correlates strongly with the actual variable: a one-standard deviation increase in the textual proxy is associated with an increase of the actual variable of about 0.2 units of standard deviation. We also compare our proxy with alternative text-based measures assuming that markets overlap with various administrative boundaries (the city, *circondario*, province, and region of the headquarters) rather than being identified based on our procedure. The correlation between the estimated market-based proxy and the actual illiteracy rate still holds when controlling for these measures.

Our contribution is closely related to three main strands of literature. The first relies on digitized newspapers to construct or complement the main variables (e.g., Ottinger and Posch (2022), Galvis et al. (2016), Sardoschau et al. (2025), Hirano and Snyder (2024), Scheve and Serlin (2023); see also Beach and Hanlon (2023b) for a comprehensive survey) or on the the use of newspaper data as instruments (see Ferrara et al., 2024). The second focuses on the usage of the presence of newspapers in a specific location to investigate the effects of information availability on a number of outcomes (e.g., Beach and Hanlon, 2023a; Gentzkow et al., 2011, 2014; Perlman and Sprick Schuster, 2016; Petrova, 2011). The third studies competition dynamics in the newspaper market and the effect of a newspaper's entry on the behavior of competitors (e.g., Bhuller et al., 2024; Cagé, 2020; Djourelova et al., 2024; Fan, 2013; Gentzkow et al., 2014; George and Waldfogel, 2006; Seamans and Zhu, 2014). We add to all these contributions by providing a more precise and granular measure of newspaper markets, improving over the often coarse proxies typically adopted in the strands of literature mentioned above.

Our work is also related to other areas of research. When investigating the geography of mass media markets, the territorial reach of radio and television stations is often inferred from official coverage reports or simulations of signal propagation (e.g., Adena et al., 2015; DellaVigna et al., 2014; Durante et al., 2019; Enikolopov et al., 2011; Olken, 2009). When these data are not available, our methodology can be adapted to such media to the extent that content can be rendered in text (through transcripts of radio or television broadcasts). Furthermore, spatial patterns of media attention may reflect historical discontinuities. In this respect, the reconstruction of newspaper markets enabled by our method speaks to the literature that exploits such discontinuities to identify causal effects (Dehdari and Gehring, 2022; Dell, 2010; Doucette, 2024; Fontana et al., 2023). Finally, a growing literature on economic history couples deep learning document layout analysis with OCR to extract structured data from historical sources (e.g., Combes et al., 2022;

[3] As detailed in Section 3, municipalities are the most granular level in the administrative hierarchy of Italy.

Correia and Luck, 2023; Dell, 2025; Dell et al., 2023; Ferguson-Cradler, 2023). We contribute to this literature by digitizing large corpora of printed historical documents and extracting relevant information from text on a scale for a period and a country for which such data are not readily available.

The remainder of the paper is organized as follows. Section 2 details the steps and characteristics of the proposed procedure. The section introduces our methodology in general terms, documenting its immediate suitability for other historical contexts, countries, languages, or data. Section 3 illustrates the context of the application of the methodology to Italian newspapers between 1919 and 1922 and provides information about the relevant historical background. Section 4 illustrates the main results of the application of our procedure to identify the markets of the considered Italian newspapers, while Section 5 compares the estimated markets with proxies based on administrative aggregations. Section 6 describes a validation exercise based on the activation of local newspapers' interest in a prominent event that spans across several localities of the country in a short period of time, the *Giro d'Italia* cycling race. Reassuringly, we find that the interest of newspapers in the race peaks in coincidence with the race crossing each specific market.[4] Finally, Section 7 concludes.

2 The methodology

This section presents all the steps of our methodology in general terms. Specific details of the application to a particular context are omitted from the discussion. We refer the reader to Section 4 for these details.

Our main goal is to reconstruct the geographic shapes of local markets for a number of newspapers at the finest feasible level of spatial granularity – at which readership data are typically unavailable – based on the textual content of the articles published by each newspaper. We describe the methodology under the assumption that newspapers are not *machine readable*, as is typical for historical sources.[5]

2.1 Digitization of textual information using *LayoutParser*

The first step requires extracting the textual information from images of potentially degraded quality. The desired output does not require any specific formatting (i.e., it is not necessary to properly identify each individual column or article), but it is essential that the number of Optical Character Recognition (OCR) mistakes are minimized to

[4]The *Giro d'Italia* is a major cycling competition that takes place across the country each year during the Spring. This setting is attractive because it provides a unique opportunity to benchmark our methodology with a common event that occurs throughout the country at a specific point in time.

[5]To process newspapers in machine-readable format, it suffices to omit the first step of the procedure and move directly to Subsection 2.2.

ensure the reliability of the subsequent steps. For this task, we build on Shen et al. (2021), combining OCR techniques with a pre-processing of the scanned textual data to maximize the readability and post-extraction processing of the candidate textual strings using a Machine Learning algorithm. More specifically, Shen et al. (2021) improve text extraction by introducing a two-stage process that first identifies the structural layout of a document and then applies OCR selectively to the identified regions. Traditional OCR engines typically process an entire page without distinguishing between text blocks, figures, or tables, which can lead to errors, especially in complex documents. By contrast, Shen et al. (2021) employ deep learning-based layout analysis to precisely delineate the 'regions' of interest, such as paragraphs or headlines, ensuring that subsequent OCR processing is applied only where appropriate. This targeted approach reduces noise from non-text elements and improves the overall accuracy and reliability of the extracted text. Crucially for our purposes, Shen et al. (2021) also pre-process the scanned text to maximize readability before any OCR is applied to digitize the text.[6]

2.2 Textual extraction of the geographic unit's name

The second step requires searching in the extracted text for references of the geographic units of interest (for example, cities). Although conceptually simple, this step comes with a number of complications and degrees of freedom that need to be taken into account.

First, the list of input names that are searched in the extracted text needs to be historically relevant at the time of newspaper publication. Suppose that one wants to search for the coverage of Oslo, Norway, in newspapers published during the early XX century. Until 1925, the city was known as Kristiania, while the current name was introduced afterward. In order to properly identify the historical coverage of today's Oslo, one must search for Kristiania for the years prior to 1925 and for the current name after 1925. City mergers or foundations must be accounted for, too. To ensure that all geographic units are searched for accurately, one needs to rely on an official list of historical names, ideally from official sources published within a short time window before the data coverage.

Second, searching for strings in a text that has been digitized with some margin of error entails a clear trade-off between type 1 and type 2 errors. A type 1 error consists in incorrectly identifying a string of text as referring to a specific city. This is a relatively frequent concern in searching for geographic units, for a variety of reasons. For example, different cities may share the same name (e.g., Portland is the name of both the largest city in Oregon and of the largest city in Maine); some cities may be named after a word of common use or, vice-versa common-use words may explicitly refer to the name of a city (e.g., the marathon running race is named after the city of Marathónas in Greece);

[6] As recommended by Shen et al. (2021), we perform the OCR using Google Cloud Vision API.

some cities may be named after famous individuals (e.g., Washington, DC). In all these occurrences – and in plenty of other examples – retrieving the name of a city in the extracted text does not ensure that the newspaper is actually referring to the city rather than to another word with the same spelling.

Type 2 errors are, instead, cases in which the extraction fails to identify a city that was mentioned in the newspaper. In an ideal scenario of perfect textual recognition and a comprehensive list of names, this should never occur. In practice, textual extraction comes with a margin of error that depends on the quality of the extraction procedure and of the input images (for example, it is possible that the extracted text contains the word *Phladelphia* instead of the word Philadelphia).

At this stage, we address the trade-off between false positives (type 1 errors) and false negatives (type 2 errors) by minimizing the former.[7] To understand why we neglect type 2 errors at this stage, it is important to note that to minimize false negatives, one should search for city names in the news content allowing for missing or misspelled letters at any place of the city name. By doing so, the risk of making type 1 errors would increase exponentially.[8] Instead, minimizing false positives (type 1 errors) has much milder consequences on the likelihood of making type 2 errors. Indeed, we will show in Subsection 2.4 that one can easily deal with false negatives at a later stage of the procedure at (virtually) no cost.

Importantly, in most cases, minimization of type 1 errors does not require simply searching for the exact name of a city. Depending on the context to which the method is applied, further refinements of the search beyond the simple name are needed. In romance languages, for example, it is useful to search for prepositions that are specific to city names before the actual name of the city, which allows one to distinguish from prepositions used when the word identifying the city is instead used under its common meaning. The needed refinements are obviously application-specific and require an appropriate knowledge of linguistic structures.

2.3 A Regression-based approach to identify cities in the market

Completing the second step allows us to measure, with a good approximation, the number of times each city i is mentioned by newspaper j in the issue published at time t. In principle, this number could be used as a proxy for the newspaper's coverage of each

[7] Note that no trade-off emerges when a portion of the full name of a city is sufficient to identify the locality unambiguously. In this case, the risk of type 2 errors can be substantially reduced by searching for that portion of the name rather than the full name, without increasing the likelihood of type 1 errors.

[8] Consider as an example the city of Rome (in Italian, *Roma*). By just altering one letter, we would identify a long list of words that are present in the Italian dictionary (such as *Coma*, *Doma*, *Toma*, *Rema*, *Rima*, *Roba*, *Rosa*) or that are used as identifiers by individuals or companies.

territory. However, it would be a very imprecise and biased proxy for the newspaper's specific interest in a locality and, in turn, for a newspaper's market. In fact, there could be a number of shocks at the newspaper level, at the publication time level, at the city level, or at the interaction of the above that may significantly affect the raw measure. A short, not comprehensive, list of specific examples of potential problems with the raw measure follows.

First, some places may happen to be covered more often than others simply because of their characteristics (e.g., large cities, cities that host famous companies or sport teams, cities that host military bases) or just because of the interaction between their characteristics and a specific period of time. For example, the Pope usually spends his summertime in the small town of Castel Gandolfo, Italy; the name of the town is usually reported in the news when the Pope is there, and journalists report his activity during the stay. The risk in those cases is to mis-interpret national salience (Calderon et al., 2023; Puglisi and Snyder, 2011) as an indicator of local interest.

Second, some newspapers may cover more cities than other newspapers simply because they are different in textual length, frequency, local vs. global focus, and ideological point of view. Importantly in historical contexts, newspapers are different in the quality of the extracted text. Moreover, each of these characteristics is not necessarily constant at the newspaper level: for example, it is not uncommon that a newspaper expands or reduces the number of pages per issue over time.

Those outlined above are examples of *confounders* that may affect the reliability of the raw count as a suitable measure of newspaper markets. Arguably, these are among the reasons why scholars prefer to approximate markets relying on administrative aggregations instead of employing textual extraction techniques. Indeed, what one is ultimately interested in when reconstructing the shape of a newspaper market is not the total number of mentions of city i in newspaper j at time t. Rather, the objective is to measure whether news about city i are published more often in newspaper j than in other newspapers or, similarly, whether newspaper j tends to cover city i more frequently than other cities. Formally, we estimate the quality of the match between city i and newspaper j, relative to the average match quality of any city with any newspaper. To this end, we construct a stacked dataset where each newspaper-city-time tuple is appended vertically. This allows us to estimate the following multi-way fixed effect specification:

$$y_{i,j,t} = \lambda_{i,t} + \psi_{j,t} + \theta_{i,j} + \varepsilon_{i,j,t}, \tag{1}$$

where the city-time fixed effect $\lambda_{i,t}$ captures all observable and unobservable characteristics of the city, potentially time-varying, which are constant across newspapers. The newspaper-time fixed effect $\psi_{j,t}$ controls for all observable and unobservable characteris-

tics of the newspaper, potentially time-varying, which are constant across cities.[9]

We focus on three dependent variables $y_{i,j,t}$: (i) the raw number of mentions of city i in the issue of newspaper j published at time t (*raw count*); (ii) the number of mentions of city i in the issue of newspaper j published at time t rescaled by the population of city i (*per inhabitant*); (iii) an indicator variable that takes value 1 if city i is mentioned at least once in the issue of newspaper j published at time t (*dummy*).

Our quantity of interest is the city-newspaper fixed effect $\theta_{i,j}$, which identifies the average match quality between city i and newspaper j, holding constant any confounders at the newspaper-time level and at the city-time level.[10] Estimating Equation (1) is computationally demanding but it allows to obtain a precise estimate of the set of city-newspaper fixed effects, which we denote $\widehat{\theta^y_{i,j}}$ (the superscript y indicates the dependent variable of each underlying regression model).[11] Importantly, $\widehat{\theta^y_{i,j}}$ is not affected by the overall importance of each city or by the length/frequency/digitization quality of each newspaper. $\widehat{\theta^y_{i,j}}$ measures the intensity to which news about city i are reported in newspaper j. We define the empirical distribution of the estimated fixed effects as Θ^y, and we identify the tails in the empirical distribution Θ^y by defining the indicator variable

$$\eta^y_{i,j} = \mathbb{1}(\widehat{\theta^y_{i,j}} > \overline{\theta^y}). \tag{2}$$

The choice of the threshold $\overline{\theta}^y$ is arbitrary: higher values ensure that fewer city-newspaper couples are flagged, so that flagged couples are likely indicating intensive coverage of city i in newspaper j – while lower values allow also less intense coverage to be flagged. We choose the value $\overline{\theta}^y$ by relying on a percentile of the distribution Θ^y that is commonly used in statistical inference, such as the 97.5 percentile (i.e., the threshold used for 5% statistical significance).[12] It is important to note that adopting a common threshold for all city-newspaper couples – instead of a different threshold for each city or for each newspaper – ensures that no restrictions are imposed on how many newspapers should flag city i or on how many cities should be flagged by newspaper j. In turn, whether more

[9] There are some analogies with the AKM model in labor economics, proposed by Abowd et al. (1999) and extensively adopted by many scholars since its first implementation. AKM models are used to identify the value-added (measured in terms of the employee's wage) due to the match quality between an employer and their employee in a regression that controls for firm fixed effects – accounting for firm specific characteristics such as the production technology – and worker fixed effects – controlling for individual characteristics such as ability or motivation.

[10] Note that the inclusion of these fixed effects is sufficient to control also for the time-unvarying characteristics of the city and of the newspaper, respectively, as well as for aggregate shocks that are constant across newspapers and city.

[11] Assuming that the data cover 50 newspapers published on average 100 times each and that the number of cities searched for is 10,000, the number of observations would be $N = 50 \times 100 \times 10,000 = 50M$ while the number of fixed effects that needs to be estimated is $K = 10,000 \times 100 + 50 \times 100 + 10,000 \times 50 = 1.5M$.

[12] Notice that when $\overline{\theta}^y$ is set at the 97.5 percentile of the empirical distribution Θ^y, the construction of the binary variable $\eta^y_{i,j}$ is analogous to a statistical hypothesis testing at the 5 percent level with bootstrapped standard errors.

or fewer cities are flagged for each newspaper depends on the density of each empirical distribution Θ_j^y relative to the density of Θ^y. If newspaper j only covers a few cities at a very high intensity, then few $\widehat{\theta_{i,j}^y}$ clear the common threshold $\overline{\theta}^y$. For other newspapers, which instead cover a higher number of cities at sufficiently high intensity, more city-newspaper couples satisfy the condition $\widehat{\theta_{i,j}^y} > \overline{\theta}^y$. That is, adopting a common threshold allows us to avoid imposing any predetermined level of concentration in the geographic structure of newspaper markets, hence obtaining fully flexible market sizes.

Different dependent variables $y_{i,j,t}$ can yield different values of $\eta_{i,j}^y$ for the same city-newspaper tuple. Since our goal is to identify cities that likely belong to the geographic area of specific coverage by newspaper j, we then define the indicator $\eta_{i,j}$ to take value 1 if all $\eta_{i,j}^y$'s are equal to 1 for all $y = \{raw\ count,\ per\ inhabitant,\ dummy\}$.

2.4 Finalization of newspaper markets using GIS techniques

The previous step of our procedure yields a preliminary list of cities that are candidates for inclusion within each newspaper's market. Yet, it is possible that preliminary markets flag some cities that are isolated from the rest of the market and leave some cities that are around those in the preliminary shape unflagged. Therefore, we finalize the shape of estimated newspaper markets leveraging spatial contiguity across localities and GIS tools to obtain coherent polygons for each estimated market.

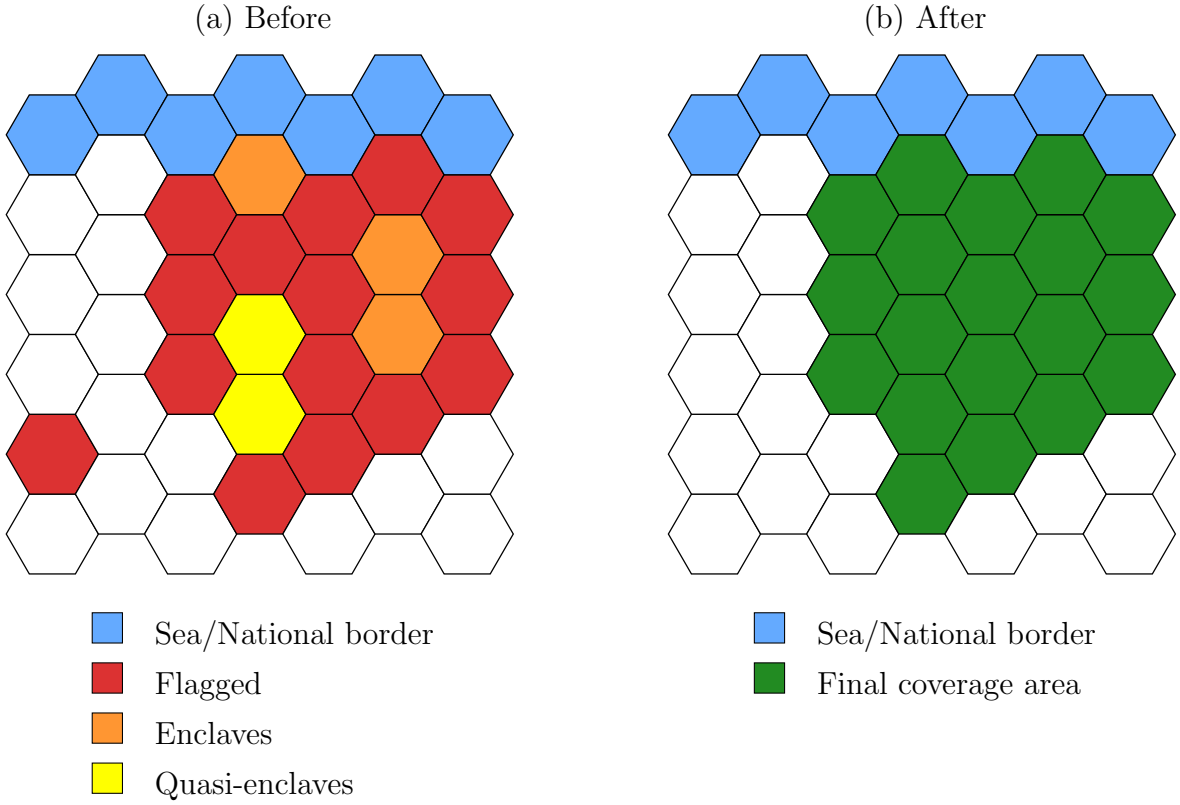
First, we remove very small clusters of flagged cities, as they are likely to be remaining type 2 errors from the step discussed in Subsection 4.2. To do so, we identify all clusters of contiguous cities included in the preliminary market of each newspaper, and exclude all clusters formed by a number of cities that is lower than an arbitrary threshold.[13] Second, we include in a newspaper's market all cities that are *enclaves* of the preliminary market.[14] Third, we smooth the external boundaries of each preliminary market. Specifically, for each newspaper, we construct a single polygon that encompasses all the cities that belong to the area at this stage. We then apply a positive buffer of N km, followed by a negative buffer of N km, obtaining a smoother polygon.[15] The final estimated market encompasses all cities that are entirely contained within this final polygon. We denote by $m(j)$ the set of localities that belong to the estimated market area of newspaper j. In turn, the

[13] National newspapers may have coverage areas featuring isolated nodes representing the major centers in the country. Our method still allows for the inclusion of such isolated locations by lowering the cluster size threshold, though this comes at the cost of producing more spatially diffuse markets for all genuinely local newspapers. The optimal choice of the threshold is context-dependent. In general, a higher threshold, resulting in more compact markets, may be preferable the further back in time the period under study goes, as slower news circulation technologies were likely to bind most newspapers to a (strictly) local dimension.

[14] We treat as *enclaves* both (i) cities entirely surrounded by flagged cities and (ii) cities bordered by flagged municipalities on all land sides, with their only interface to non-flagged territory being the coast or the national border.

[15] We denote the cities that are included in the estimated market only in this final stage as the market's *quasi-enclaves*.

Figure 1: Reconstruction of a newspaper's market with GIS tools



Notes: The figure illustrates the steps of the GIS-based filtering procedure. Each hexagon represents a city. Panel (a) shows flagged cities in red, enclaves in orange, and quasi-enclaves in yellow. Blue-colored cells represent the national border (or sea). Panel (b) displays the final estimated market in green.

binary indicator $m_{i,j}$ takes value 1 if $i \in m(j)$.

Figure 1 illustrates the implementation of these steps if we require clusters to contain at least two cities. Panel (a) displays two different enclave clusters of cities, in orange, and a quasi-enclave cluster, in yellow. The three clusters are included in the final market, reported in Panel (b), as a result of the GIS corrections. In contrast, the isolated flagged city on the bottom left of the map is excluded from the final market as it does not belong to a cluster of at least two contiguous cities.

The GIS procedure outlined in this section has an important implication for the robustness of our estimated markets: it is neither sufficient nor necessary for a city to be flagged in the regression-based approach in order to be included in the market. On the one hand, if a city is flagged but its neighbors are not, we exclude it from the market. In fact, it is likely the city has been flagged due to idiosyncratic shocks that our multi-way fixed effect could not capture or due to type 2 errors in the processing of textual information. On the other hand, if a city is not flagged but all its neighbors are, the city is included in the final market. Indeed, it is likely that such city would have been covered extensively by the newspaper had some important event taken place there. It is

important to note that our GIS filtering procedure does not rely on any administrative aggregation (it is necessary and sufficient to observe the shapes of the units at which the previous steps are performed) or distance measures from a newspaper's headquarters. More precisely, our methodology treats the city in which a newspaper is headquartered exactly as any other city; for this reason, headquarters do not necessarily represent the focal points of the estimated markets.

3 Historical background

In the remainder of the paper, we apply the procedure illustrated in Section 2 to newspapers published in Italy in the period 1919–1922. Documenting that the proposed methodology works well in a turbulent historical context is informative both about the overall validity of the procedure and its applicability to other historical contexts. This section provides a concise description of the relevant historical background.

Newspapers: readership and diffusion. In the early twentieth century, Italy was characterized by a rapidly growing literacy rate. Every year many newspapers were founded, and others ceased their operations. Several linguistic minorities, with their own newspapers, were (and still are today) present close to the borders with France and Austria, allowing us to test our methodology on multiple languages. Furthermore, the political instability characterizing the years between the end of WWI in November 1918 and the rise to power of the Fascist Party in October 1922 provided a fertile background for the publication of very diverse newspapers (many of which available only for a portion of the considered period).

Compared to other European countries and the United States of America, the number and the diffusion of printed newspapers in Italy had been limited during the XIX century by the high illiteracy rates of the population: 78 percent of the population was illiterate when the country was unified in 1861, as compared to approximately 20 percent in Germany and the U.S. and 47 percent in France. In the following decades, the rate of illiteracy converged to significantly lower values, easing the granular diffusion of newspapers. By the 1921 Census, the illiterate population dropped to 29 percent. In turn, newspapers established themselves as the main source of information in the early XX century and at least until the foundation of the public radio broadcaster URI in October 1924.[16] Reconstructing precisely the number of newspapers available in Italy in the early XX century is difficult, as the official historical statistics report a fragmented pattern. According to Ministero di Agricoltura, Industria e Commercio (1908), a total of 1,901 periodicals were published in Italy in 1895. This number includes many pamphlets that would hardly be considered newspapers according to contemporary standards. Only 45

[16]Prime Minister Giovanni Giolitti declared in 1913 that Italians were consuming approximately 5 million newspapers on a daily basis, which amounts to 0.2 newspapers per inhabitant.

percent of such pamphlets were published on a high frequency basis (i.e., at least once per week) and only 27 percent of them covered political events.[17] Moreover, many of the publications included in the official statistics were short lived: 57 percent of the publications available in 1895 ceased their operations during the same year, and 50 percent of the publications available in 1895 were founded during the same year. Therefore, the number of effective newspapers available in Italy at the time appears much smaller than the figures reported in the official statistics. According to CENSIS (2004), only 66 daily newspapers were published in Italy during the interwar period.

The administrative hierarchy of Italy in 1921. Municipalities were (and still are) the most granular level of the administrative hierarchy. However, the number and geographic boundaries of municipalities have changed significantly over time as a result of several municipal mergers and splits. Moreover, Italy lost part of its territory on the North-East border – to current Croatia and Slovenia – after the end of WWII. In the 1921 census, Italy counted 9,195 municipalities, of which 9,019 belong to contemporary boundaries. Being the lowest level of administration, municipalities in Italy were small: the average population was approximately 4,300 inhabitants and the median population approximately 2,100 inhabitants, while the average surface was 3,400 hectares.

Above municipalities, there were three additional administrative aggregations (from large to small): regions (18), provinces (75), and *circondari*. Regions, comparable in surface to U.S. commuting zones, encompassed an average of 511 municipalities (with approximately 2.2M inhabitants in the 1921 census).[18] Provinces were nested within regions and their surface is comparable to that of U.S. counties (with an average size of 123 municipalities and 0.5M inhabitants, approximately). *Circondari* were nested within the provinces and each *circondario* was formed by a cluster of neighboring municipalities.[19] In 1921, Italy counted 245 *circondari*, on average containing 38 municipalities each (with about 0.16M inhabitants).

4 The estimation of local markets

In this section, we discuss and present the application of our procedure to a sample of 154 daily or weekly newspapers available in Italy during the period 1919–1922. For brevity,

[17] Official statistics do not provide information on the number of daily or weekly publications that covered political events.

[18] Regions were formally established only after WW2, even though their unofficial boundaries had been used since the Unification as statistical districts in population censuses.

[19] Another administrative aggregation, the *mandamento*, was mainly used for judiciary purposes. *Mandamenti* were not considered as a statistical unit of observation, which implies that information on their geography is not available. It should be noted that the same municipality could be divided into several *mandamenti*. When comparing our estimated market boundaries against administrative aggregations, we focus on *circondari*, provinces, and regions since the shapefiles of the distribution of *mandamenti* are not available.

we omit many of the general details already illustrated in Section 2 and focus instead on the specific application of the procedure to our data. In what follows, i indicates the municipality, j the newspaper, and t the week of publication of the issue. The procedure allows us to identify whether each municipality belongs to the market $m(j)$ of newspaper j.

4.1 Newspapers in the sample

We collected scanned copies of 154 weekly or daily newspapers published in Italy between 1919 and 1922, for a total of more than 60,000 issues and approximately 300,000 pages. Due to the impossibility of retrieving reliable information on the overall number of newspapers published in Italy at the time, we cannot establish how close to the universe is the number of newspapers we retrieved, and whether our sample is representative of the news content available in the period. Figure A.1 in Appendix A reports the main characteristics of the newspapers sampled.

Some of the newspapers in our sample are still active today: among them, established newspapers with national diffusion (e.g., *La Stampa, Il Messaggero, Il Sole 24 Ore*) and several local newspapers (e.g., *La Gazzetta d'Alba*, founded in 1882 in the Piedmontese town of Alba and still active). As other languages are commonly used in regions close to international borders, our sample also includes Italian newspapers written in a foreign language.[20] Some newspapers have a strong ideological leaning, often immediately conveyed by the newspaper's title; others use the title to signal their local focus. Specifically, 28 newspapers mention their ideology in name, and 74 of them make an explicit reference to a city or area in Italy.[21] 46 newspapers are published daily, while the remaining ones are published weekly or twice per week. 77 newspapers span the entire four-year period while the remaining ones are available for a shorter portion of the period only. Figure A.2 in Appendix A reports the number of newspapers in our dataset for each week between the first week of 1919 and the last week of 1922. The figure documents increasing numbers between 1919 and 1921: the average number of newspapers in our sample is approximately 90 in 1919 and 120 in 1921. It also shows a reduction in the availability of newspapers during 1922, when the Fascist movement attacked numerous left-leaning newspapers before and after their rise to power.

Figure 2 plots the geographic distribution of the headquarters of the newspapers in our sample.[22] Many of the newspapers in our sample are published in Northern Italy. This

[20] We have data on four newspapers published in French, concentrated in the North-Western region of Aosta Valley, and eight newspapers published in German in the South-Tyrolean territories annexed by Italy at the end of World War I.

[21] Our sample covers multiple newspapers per each of the political ideologies represented in the Italian Parliament before the establishment of the Fascist regime (Communists, Socialists, Catholic, Liberal-Democratic, Republican, Nationalist, and Fascist).

[22] When more than one newspaper is published at the same location, the size of the circle in the figure

Figure 2: Distribution of newspaper headquarters

Notes: The map reports the location of newspaper headquarters for the newspapers included in our sample. The size of each red marker is proportional to the number of headquarters located in the same municipality. The maximum value is Rome (where the headquarters of eight newspapers are located). Region boundaries as of 1921.

should not come as a surprise due to the regional differences in economic development, literacy rates, and population density between the North and the South of the country. In Table A.1 in Appendix A, we show that the geographic distribution of a newspaper's headquarters strongly correlates with each of those characteristics. Moreover, Figure A.3 in Appendix A documents a very strong correlation between the geographic distribution of newspaper headquarters in our sample and that of headquarters of the universe of printed publications in 1905, as reported in Ministero di Agricoltura, Industria e Commercio (1908). We conclude that the observed geographical variation in our sample coverage does not reflect systematic issues concerning the representativeness of our sample.

reflects the number of newspapers.

4.2 The textual search for the name of municipalities

We digitize all scanned copies using Shen et al. (2021)'s *Layout Parser*.[23] In order to search for each municipality in the extracted text, we use the official municipality name as of the 1921 Census according to the Italian Institute of Statistics (ISTAT). For municipalities located in areas of Italy where other languages are widely spoken among the population (that is, French in the Aosta Valley and German in *Alto Adige*/South Tyrol), we search for both the Italian version of the municipality name and the foreign version of the municipality name. We manually inspect all occurrences of municipality names that are formed by more than one word and restrict the search to the first word only if it is sufficient to identify the municipality unambiguously. Moreover, we also search for widely used alternatives to the official names of some municipalities (e.g., *Reggio Calabria* for *Reggio di Calabria*), and abbreviated forms (e.g., *St. Ulrich* for *Sankt Ulrich*, the German toponym of *Ortisei* in South Tyrol). To minimize the occurrence of false positives if a municipality's name is a word of common use in Italian, we search for a string containing one of the prepositions used in Italian to introduce the name of a locality followed by a single space and the name of the municipality (or its first word, whenever feasible).[24] We are unable to unambiguously identify only 109 municipalities (i.e., 1.2% of the sample) based on the municipality's name because of some homonymy with another municipality. In those occurrences, we assign matches with the newspaper's content to both municipalities if we identify the string as explained above.[25] Based on the output of the search of municipality names preceded by a proposition, 95 percent of municipalities (equivalent to 97 percent of the 1921 Census population) are mentioned at least in one newspaper.[26]

[23] Eighteen newspapers in our sample are also available in searchable PDF format, as their text content was previously digitized using standard OCR routines by libraries and archival institutions. However, the quality of the extracted text obtained using our approach – based on *Layout Parser* and Google Cloud Vision – is substantially improved. When we apply the same text extraction and processing procedure described later in this subsection to both versions of the digitized content of these 18 newspapers, we identify 4,224 unique municipalities in our re-digitized texts (with an average of 62.11 mentions per municipality), compared to 3,722 municipalities identified in the pre-digitized versions (with an average of 48.02 mentions). Focusing on the four largest Italian cities in 1921, Milan is mentioned 19,083 times in our digitized version and 13,888 times in the existing one; Rome 14,357 vs 9,550; Turin 8,584 vs 5,257; and Naples 3,159 vs 2,256.

[24] Specifically, the prepositions are *a, da, di, per, presso,* and *verso*. Note that we do the same for newspapers written in French and German, searching for the corresponding versions of the relevant prepositions: *à, en, de, chez, vers, pour* for French, and *aus, von, nach, zu, in, bei* for German.

[25] There is only one case of homonymy between three municipalities. In that occurrence, we assign matches with the newspaper content to the three municipalities if we identify the string as explained above.

[26] At this stage of the procedure we slightly aggregate municipality-level information because geographic information on municipal boundaries is not available for the years prior to 1991. Specifically, using administrative data from ISTAT on all municipal mergers and splits that occurred over time, we construct a crosswalk between municipalities in the 1921 census and municipalities in the 1991 census. In the case of a merger, the spatial unit of observation is defined as the area of the 1991 municipality, and the number of mentions is calculated as the sum of mentions across all municipalities involved in the merger in each issue of newspaper j published at time t. In the case of splits, each newly established municipality (i.e., those established between 1921 and 1991) is traced back to the preexisting municipality from which

Not surprisingly, the city of Rome is the most cited (with about 349,000 occurrences in 153 different newspapers), while 708 municipalities are cited by only one newspaper each. As documented in Figure A.4 in Appendix A, the average newspaper mentions 809 municipalities. The newspaper that mentions fewer municipalities (22) is a local pro-socialist pamphlet published in the southern city of Potenza, while the newspaper that mentions the largest number of municipalities (more than 3,600) covers almost half of them during the four-year period.[27]

The distribution of mentions across municipalities confirms that such rough measures can not be a reliable proxy for the geography of newspaper markets. This does not necessarily imply that many of the municipalities that are mentioned are of specific interest to a newspaper. Looking at the municipalities that are mentioned by each newspaper, it becomes apparent that textual information has some predictive power but comes with several false positives and false negative. As an example, Figure 3 illustrates the three raw measures across all issues of the local newspaper *La Provincia Pavese*, published in the city of Pavia, Lombardy. Panel (a) shows the raw count of mentions by municipality, while Panel (b) rescales this count by the municipal population. Finally, Panel (c) displays a binary dummy that takes value 1 for all municipalities that are mentioned at least once. All three maps reveal a consistent spatial pattern, with a clearly defined cluster of frequently mentioned municipalities concentrated in the North-West of the country, exactly in the area around the headquarters of the *La Provincia Pavese*. At the same time, mentions are observed, with varying frequencies, across a broader set of locations throughout the country. All metropolitan areas of the country are flagged, although many of them are unlikely to belong to the newspaper's area of interest. The next subsections build increasingly more refined market shapes for the *La Provincia Pavese*, making it easy to visualize how each step contributes to the estimation of final markets.

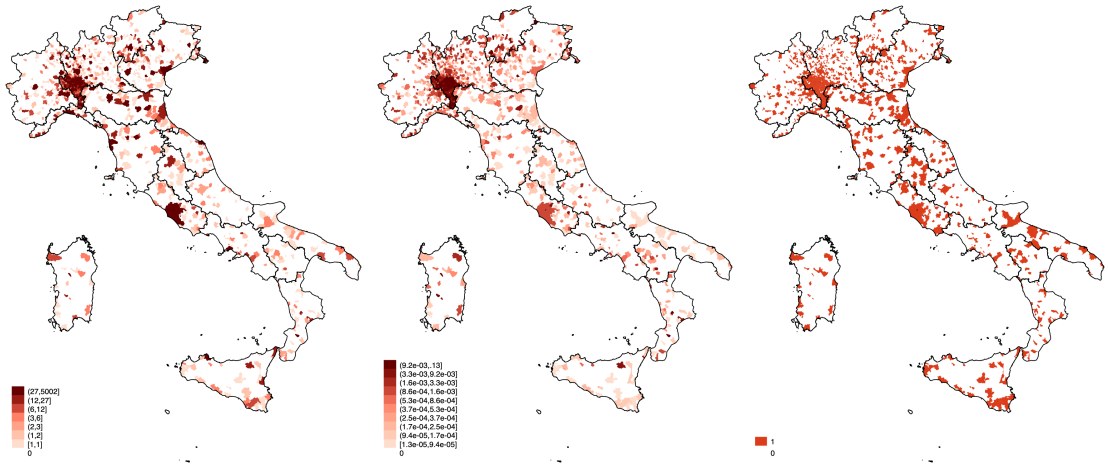
4.3 The estimation of preliminary markets

We construct a three-dimensional stacked panel as explained in Section 2. To ensure comparability between daily newspapers and weekly newspapers, we aggregate the number of mentions of municipality i across all issues of newspaper j published in week t (so that the data time dimension is the week). In doing so, we obtain a panel with approx-

it was detached. In this case, the spatial unit of observation corresponds to the combined area of the original municipality and of all municipalities that were created by splitting the original one, and the number of mentions is defined as that of the corresponding 1921 municipality. This procedure yields a list of 7,921 "pseudo-municipalities" based on 1921 data mapped to 1991 boundaries, which we use as units of observation in later steps. This adaptation is largely inconsequential.

[27] Table A.2 presents the descriptive statistics of the three raw measures of coverage of municipality i in newspaper j at time t, while Table A.3 shows that the three variables are highly correlated with each other.

Figure 3: *La Provincia Pavese* – Municipality mentions



(a) raw count (b) per inhabitant (c) dummy

Notes: Region boundaries as of 1921 and municipality boundaries as of 1991. See Footnote 26 for details.

imately 180M observations.[28] Estimating Equation (1) requires the estimation of three separate regression models, each of which having just less than 3 million fixed effects and predicting 1.22M values for $\widehat{\theta^y_{i,j}}$.[29]

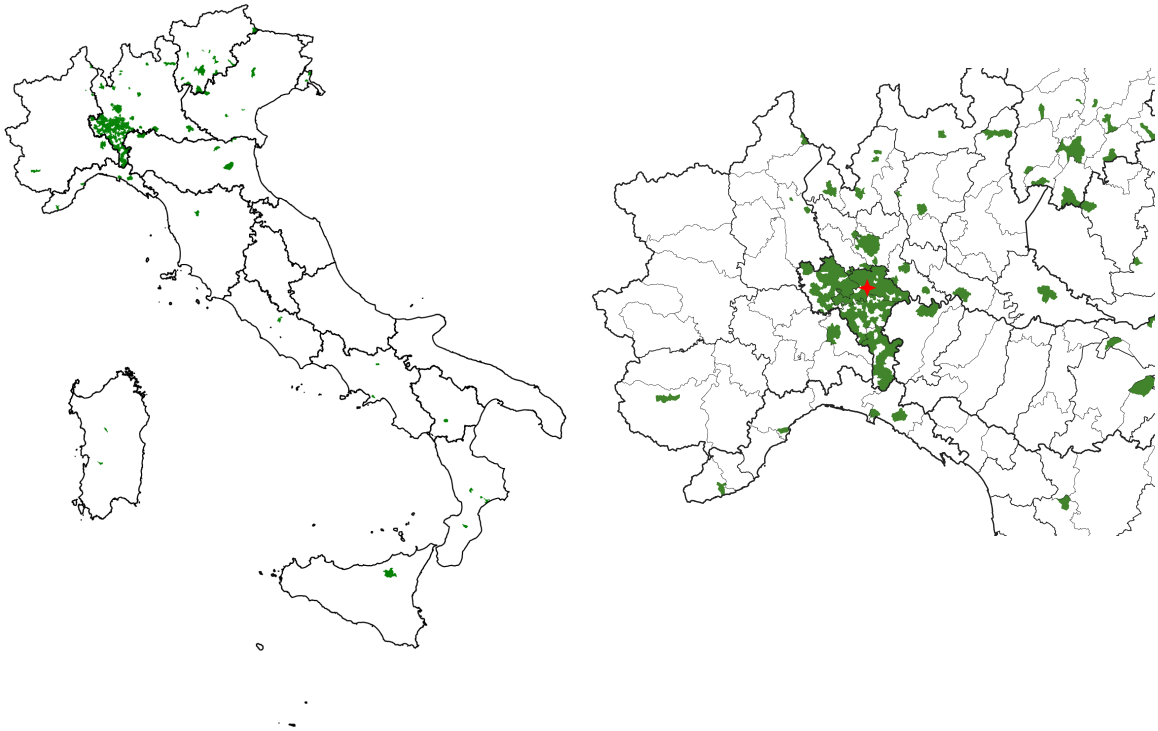
We denote by Θ^y the empirical distributions of $\widehat{\theta^y_{i,j}}$. Such distributions are characterized by extreme skewness and kurtosis (see Table A.4 in Appendix A). Intuitively, many municipality-newspaper couples display low values of $\widehat{\theta^y_{i,j}}$ because municipality i does not belong to the area of interest targeted by newspaper j. If, instead, $\widehat{\theta^y_{i,j}}$ is high, which occurs infrequently, the data suggest that newspaper j is particularly interested in reporting news about municipality i. Figure A.5 in Appendix A, in which all values of $\widehat{\theta^y_{i,j}}$ are reported in ascending order with the exclusion of extreme outliers, illustrates the distributions. The red vertical line denotes the threshold $\overline{\theta^y}$ placed at the 97.5th percentile of each distribution Θ^y. Relying on Condition (2), we assign value $\eta^y_{i,j} = 1$ to municipality-newspaper pairs for which $\widehat{\theta^y_{i,j}}$ is larger than $\overline{\theta^y}$. The number of municipality-newspaper pairs that clear the threshold is fixed in the full sample. However, depending on the shape of each newspaper-specific distribution Θ^y_j, the number of pairs that are flagged (i.e., such that $\eta^y_{i,j} = 1$) may vary substantially between newspapers.

Figure A.6 in Appendix A shows the variation between newspapers in the number of municipalities i that clear the threshold $\overline{\theta^y}$ for each newspaper j and each dependent variable y. More specifically, each panel reports the distribution of the quantity $\frac{\sum_{i=1}^{N} \eta^y_{i,j}}{N}$

[28] Each newspaper-week tuple appears in the data only if newspaper j was published (and digitized) at least once in week t. Each tuple is then multiplied by the total number of (pseudo-)municipalities. If municipality i is not mentioned in any of newspaper's j issues published in week t – but the newspaper has been digitized for the same week, we manually assign value 0 to the related triple.

[29] The results of the prediction exercise seem accurate, with adjusted R^2 of 0.75 for the raw count, 0.18 for the count per inhabitant, and 0.37 for the dummy, respectively.

Figure 4: *La Provincia Pavese* – Preliminary market



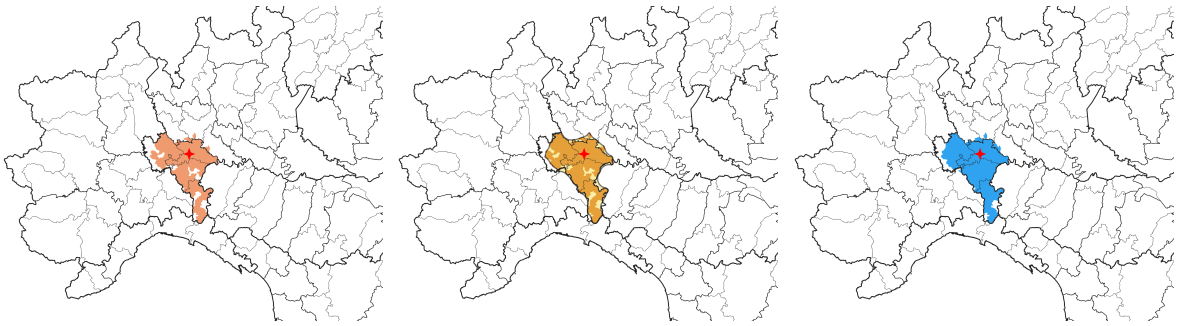
Notes: Region, province, and *circondario* boundaries as of 1921 and municipality boundaries as of 1991. See Footnote 26 for details. The red marker indicates the location of the newspaper's headquarters.

calculated separately for each of the 154 newspapers in our sample. The final output of the regression stage is presented in Figure A.7 in Appendix A, in which we report the percentage of municipalities included in the *preliminary* market of each newspaper. Figure 4 reports the set of flagged municipalities for the *La Provincia Pavese*, showing that our multi-way regression model identifies significantly 'cleaner' preliminary markets compared to those based on the raw frequency of mentions. Yet, as highlighted by the left panel of the figure depicting the whole country, a non-negligible number of municipalities far from the city of Pavia are included, oftentimes as singleton clusters. At the same time, as shown on the right panel depicting only the North-West of the country, some municipalities in the inner part of the preliminary market are not flagged.

4.4 Isolated clusters, enclaves, and boundary adjustment

To implement the filtering procedure outlined in Subsection 2.4 (aimed at obtaining coherent shapes based on municipalities flagged by the regression-based approach), we construct an adjacency matrix for all Italian municipalities as of 1991. We impose a minimum cluster size threshold of five flagged municipalities, hence excluding all clusters that fall below this cutoff. This threshold is rather conservative, ensuring that the resulting markets are relatively compact. Figure 5 outlines the procedure focusing on the

Figure 5: *La Provincia Pavese* – reconstruction of shapes using GIS



(a) Clusters of ≥ five municipalities

(b) Smoothed polygon

(c) Final estimated market

Notes: Detail of North-Western Italy. Region, province, and *circondario* boundaries as of 1921 and municipality boundaries as of 1991. See Footnote 26 for details. The red marker indicates the location of the newspaper's headquarters.

La Provincia Pavese. We exploit the 1991 adjacency matrix to exclude all clusters that are smaller than the specified threshold. We then incorporate all enclaves into the list of municipalities to include in the market. The resulting intermediate output is shown in Panel (a). Although the shape of the preliminary market appears to be already well defined, a number of municipalities remain excluded despite being almost entirely surrounded by flagged ones. In order to address the issue, we apply the buffer-debuffer routine that identifies and includes quasi-enclaves. Panel (b) of Figure 5 shows the resulting smoothed shape. Finally, we include all municipalities the surface of which is entirely contained within the polygon. The final estimated market, illustrated in Panel (c), does not perfectly overlap with the boundaries of the province of Pavia (where the newspaper's headquarters were located), nor does it overlap with Pavia's *circondario* or the Lombardy region. On the one hand, some municipalities in the northern part of the market belong to the neighboring province of Milan. On the other hand, several municipalities within the province of Pavia do not belong to the estimated market. Interestingly, most of the non-included municipalities are located in a part of the province, corresponding to the area of *Lomellina*, that followed a different historical trajectory than the rest of the province. While the area of Pavia remained under the Habsburg's rule along with the Duchy of Milan until 1859, *Lomellina* had been part of the Kingdom of Sardinia since the XVIII century. The *Lomellina* joined the city of Pavia into a unique province only during the period of Italian unification. Our methodology can uncover such historically rooted distinctions, capturing the underlying territorial focus of media outlets. This is a feature that proxies for newspaper markets based on administrative aggregations are missing, as they disregard historical roots that continue to affect local communities over time (e.g., Dehdari and Gehring, 2022; Dell, 2010; Doucette, 2024; Fontana et al., 2023).

Figure 6 shows the estimated markets of all newspapers on a single map of Italy. As

Figure 6: Visualization of estimated newspaper markets

Notes: Region and province boundaries as of 1921 and municipality boundaries as of 1991. See Footnote 26 for details.

discussed in Section 3, our sample is geographically skewed toward the Center-North of the country. Many areas in the South are not covered by any of the newspapers in our dataset, and the two major islands – Sicily and Sardinia – are covered by a single newspaper each. Overall, 58 percent of municipalities belong to at least one market. Among them, 44 percent (corresponding to 25.5 percent of the total number of municipalities) are covered by only one newspaper, while 6.9 percent (4 percent of the total) belong to the market of five or more newspapers. The share of municipalities included in each newspaper's *final* market is illustrated in Figure A.8 in Appendix A.[30] The size of final markets ranges from 0.08 percent (6 municipalities) to 8.5 percent of municipalities (670 municipalities). The median and average market sizes are 0.57 percent (45 municipalities) and 0.93 percent (73 municipalities), respectively. Only seven of the 154 newspapers in the sample do not have a well-defined market that satisfies our criteria. Six of them are published by local branches of national political organizations, and their content tends to be more similar to that of propaganda leaflets than to local news reporting.[31] The seventh newspaper, *il*

[30]Figure A.9 in Appendix A illustrates the positive and significant correlation between the size of the *final* markets and the size of the *preliminary* markets.

[31]None of their estimated city-newspaper fixed effects satisfies Condition (2) for all model specifications – i.e., there is no municipality i such that $\eta_{i,j}^{y} = 1$ for all y. The names of the outlets and the locations

Popolo d'Italia (founded and directed by Benito Mussolini) lacks a clearly defined market as a result of the GIS-based correction for isolated clusters and enclaves as, despite being based in Milan, focused exclusively on national political affairs.

5 The advantages of our methodology vs. proxies based on administrative aggregations

Our estimation methodology yields substantially different newspaper market shapes than those obtained based on administrative aggregations. We show that only a small share of estimated markets can be properly proxied by administrative units. Following Snyder and Strömberg (2010), we compute a congruence score between each estimated market and an administrative unit defined as

$$Congruence^{Pop}_{m(j),a(j)} = \frac{Pop_{m(j)}}{Pop_{a(j)}},$$

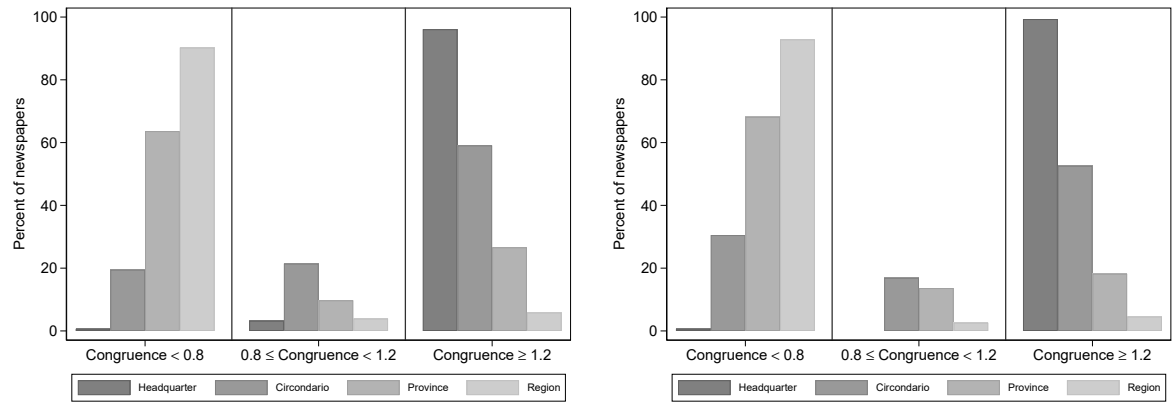
where $Pop_{m(j)}$ is the total population (1921 census) living within the estimated media market $m(j)$, while $Pop_{a(j)}$ is the total population living within the administrative aggregation $a(j)$ including the headquarters of newspaper j. We compute this score for four administrative levels containing newspaper headquarters in 1920s Italy, ordered from the most to the least granular: (i) city; (ii) *circondario*; (iii) province; and (iv) region. The congruence scores capture how well each administrative aggregation proxies for the shapes of the estimated markets. We also replicate the exercise using a surface-based congruence score, replacing the population with the land surface, that is,

$$Congruence^{Sur}_{m(j),a(j)} = \frac{Sur_{m(j)}}{Sur_{a(j)}},$$

where $Sur_{m(j)}$ (resp., $Sur_{a(j)}$) is the total area (in hectares) within the boundaries of the estimated media market (resp., within the administrative aggregation) according to the census of 1921. Values of the congruence scores close to one indicate a good approximation, whereas values away from one suggest that the shapes are significantly different. Figure 7 shows the percentage of newspapers that experience values of the congruence score close to one (respectively, away from one). Low-congruence newspapers are those that have a congruence score below 0.8 or above 1.2, implying a discrepancy between the shapes that exceeds 20 percent. As shown in Figure 7, only a small fraction of newspaper markets are well approximated – i.e., have a score between 0.8 and 1.2 – by any administrative-boundary proxy, both for the population-based measure (Panel (a))

of their headquarters are: *Avvenire Anarchico* from Pisa, *Bandiera Bianca* from Udine, *Giovinezza* and *Italia Antibolscevica* from Milano, *Lotte Civili* from Piombino, *Popolo Pistoiese* from Pistoia.

Figure 7: Congruence between estimated markets and proxies based on administrative aggregations



(a) Population

(b) Surface

Notes: Each bar reports the percent of newspapers such that the congruence between estimated markets and proxies based on administrative aggregations is below 0.8; between 0.8 and 1.2; above 1.2, respectively. In Panel (a), we calculate congruence based on the population of each municipality as of the 1921 Census. In Panel (b), we calculate congruence based on the surface of each municipality as of the 1921 Census.

and for the surface-based measure (Panel (b)). Larger administrative aggregations tend to overestimate market shapes, while lower aggregations systematically underestimate newspaper markets.

We further illustrate the lack of correspondence between estimated markets and administrative aggregations in Figures A.10 and A.11 in Appendix A. In Figure A.10 (respectively, A.11), the left panel reports, for each newspaper, the share of the estimated market's population (resp., surface) that lies within the headquarters' administrative aggregation. The right panel reports the share of the population (resp., land surface) of the headquarters' administrative aggregation that falls outside the estimated market.[32] The two figures highlight substantial variation across newspapers for each administrative aggregation, with no aggregation systematically outperforming the others (none of them consistently yields values close to one in the left panel and close to zero in the right panel). Together, Figures 7, A.10, and A.11 document substantial differences between the results of our estimation and those based on administrative aggregations.

To assess whether our estimated markets, besides being different, outperform the proxies usually adopted in the literature, we compare proxies of illiteracy obtained from the textual content of newspapers with the official census rates of illiteracy – available for 1921 at the municipality level. More precisely, we show that such text-based measures

[32]Formally, we show the ratio $\frac{Pop_{m(j)\cap a(j)}}{Pop_{m(j)}}$ and $\frac{Sur_{m(j)\cap a(j)}}{Sur_{m(j)}}$ in the left panels of the two figures, and the ratio $\frac{Pop_{a(j)}-Pop_{m(j)\cap a(j)}}{Pop_{a(j)}}$ and $\frac{Sur_{a(j)}-Sur_{m(j)\cap a(j)}}{Sur_{a(j)}}$ in the right panels, where $a(j)$ stands for headquarters' city, *circondario*, province, and region, respectively.

provide a more precise proxy of local economic conditions when the textual content is projected onto municipalities using estimated shapes (based on our methodology) rather than administrative aggregations. Specifically, we parse the digitized textual content of each newspaper to identify mentions of (il)literacy-related keywords in each issue.[33] Formally, we define the index $Illiteracy_j$ as

$$Illiteracy_j = \frac{1}{No.Issues_j} \sum_{k(j)=1}^{No.Issues_j} \frac{MentionsIlliteracy_{k(j)}}{No.Pages_{k(j)}} ,$$

where $MentionsIlliteracy_{k(j)}$ is the raw number of mentions of illiteracy-related keywords in issue k of newspaper j, $No.Pages_{k(j)}$ is the number of pages of text in issue k of newspaper j, and $No.Issues_j$ is the total number of issues of newspaper j that we observe. The index $Illiteracy_j$ adjusts the raw number of mentions for the length of the issue, the frequency of publication, and the window of observation, ensuring comparability across newspapers.

The final step is to project $Illiteracy_j$ onto municipalities. For each municipality i, we calculate the weighted average of the $Illiteracy_j$ scores for all newspapers whose estimated markets encompass the municipality, where the weights are proportional to the share of the population of market $m(j)$ that lives in municipality i. Formally, we define

$$Illiteracy_i^{Pop} = \frac{\sum_{m(j)=1}^{M} Illiteracy_j \frac{Pop_i}{Pop_{m(j)}} \times \mathbb{1}[i \in m(j)]}{\sum_{m(j)=1}^{M} \frac{Pop_i}{Pop_{m(j)}} \times \mathbb{1}[i \in m(j)]} , \tag{3}$$

where $\frac{Pop_i}{Pop_{m(j)}}$ is the ratio between the population of municipality i and the total population in the newspaper market $m(j)$, and M denotes the set of all $m(j)$. Similarly, we use the municipality surface to calculate the index $Illiteracy_i^{Sur}$ with weights proportional to the share of the municipality surface relative to the total surface of each market.

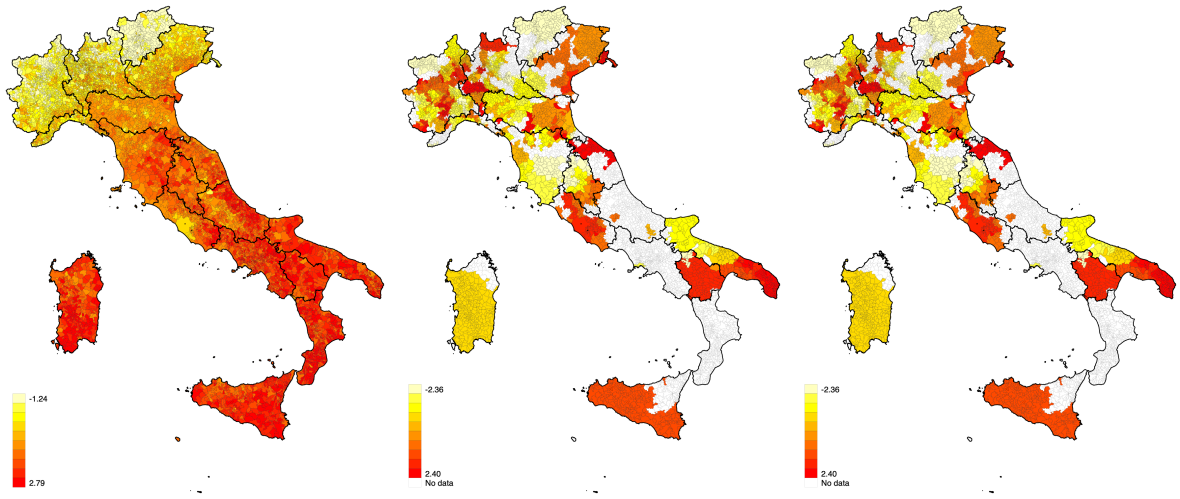
Figure 8 compares the spatial distribution of actual illiteracy rates according to the 1921 census (in Panel (a)) and the text-based proxies $Illiteracy_i^{Pop}$ and $Illiteracy_i^{Sur}$ (in Panels (b) and (c)).[34] Although our newspaper sample has limited coverage of some southern regions, the distribution of journals' interest exhibits a clear North-South gradient that mirrors the census pattern. Looking at the North of the country, a West-East gradient also emerges: newspapers based in Piedmont (in the North-West) pay less attention to illiteracy than newspapers covering areas in the North-East.[35] A formal comparison of the variables indicates that a one-standard deviation increase in either $Illiteracy_i^{Pop}$ or

[33] The list comprises the Italian terms: *analfabeta/i*, *(an)alfabetismo*, *alfabetizzazione*, *alfabetizzato*, *istruzione*, *educazione*, *lettura*, *lettore/i*, *scuola* – equivalent to (il)literate(s), (il)literacy, education, reading, reader(s), school.

[34] We standardize all variables to have mean zero and unit variance.

[35] Note that Piedmont was the region experiencing the lowest illiteracy rate of the country.

Figure 8: Illiteracy rate –actual data and textual extraction



(a) $Illiteracy_i$ (b) $Illiteracy_i^{Pop}$ (c) $Illiteracy_i^{Sur}$

Notes: In all panels, illiteracy measures are standardized to have mean 0 and standard deviation 1. Each color represents one decile in the empirical distribution, with darker colors capturing higher illiteracy rates (Panel (a)) or higher intensity of illiteracy-related keywords in the newspapers' texts (Panel (b) and Panel (c)), respectively.

$Illiteracy_i^{Sur}$ predicts an increase in the actual illiteracy rate of approximately 20 percent of a standard deviation.

Figure A.12 in Appendix A shows maps that document the geographical distribution of the text-based proxy for the illiteracy rate calculated using administrative aggregations (region, province, *circondario*, and city hosting a newspaper's headquarters).[36] The spatial patterns broadly resemble Panels (b) and (c) of Figure 8 but the set of municipalities for which the measure is defined varies substantially across administrative levels. As documented in Figure A.13 in Appendix A, the number of municipalities with a non-missing proxy for illiteracy obtained following our approach is similar to that of province-based projections, smaller than region-based projections, and larger than *circondario* and city-based projections. Furthermore, all measures are significantly different from those reported in Figure 8 for municipalities sufficiently far from newspaper headquarters.

To assess whether estimated newspaper markets provide additional information compared to those based on administrative aggregations, we estimate linear models that regress the actual illiteracy rate (1921 census) on $Illiteracy_i^{Pop}$, while controlling for the

[36] Formally, for each municipality i and each administrative aggregation a, we compute $Illiteracy_i^{Pop,a} = \frac{\sum_{a(j)=1}^{A} Illiteracy_j \frac{Pop_i}{Pop_{a(j)}} \times \mathbb{1}[i \in a(j)]}{\sum_{a(j)=1}^{A} \frac{Pop_i}{Pop_{a(j)}} \times \mathbb{1}[i \in a(j)]}$ and, analogously, $Illiteracy_i^{Sur,a} = \frac{\sum_{a(j)=1}^{A} Illiteracy_j \frac{Sur_i}{Sur_{a(j)}} \times \mathbb{1}[i \in a(j)]}{\sum_{a(j)=1}^{A} \frac{Sur_i}{Sur_{a(j)}} \times \mathbb{1}[i \in a(j)]}$.

Table 1: Correlation between actual illiteracy rate and textual extraction (population)

	(1)	(2)	(3)	(4)	(5)
Dep. var.:	Illiteracy rate (1921 census)				
Illit. newsp. incid.	0.203***	0.0722***	0.0911***	0.155***	0.537***
	(0.0119)	(0.0155)	(0.0193)	(0.0293)	(0.144)
Observations	4,594	4,515	3,970	2,876	76
R^2	0.049	0.094	0.102	0.116	0.252
Region text-based illiteracy		✓	✓	✓	✓
Province text-based illiteracy			✓	✓	✓
Circondario text-based illiteracy				✓	✓
Headquarters text-based illiteracy					✓
Mean dep. var.	–0.243	–0.255	–0.312	–0.492	–0.553

Notes: The unit of observation is a municipality (1991 boundaries, see Footnote 26 for details). Column (1) reports the correlation between the standardized illiteracy rate (1921 census) and the standardized incidence of illiteracy-related keywords, projected onto municipalities using estimated markets. Columns (2)–(5) augment the specification with controls for the standardized incidence of illiteracy-related keywords projected onto municipalities using region boundaries, province boundaries, *circondario* boundaries, and headquarters' city boundaries, respectively. Standard errors robust to heteroskedasticity are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

text-based proxies calculated using administrative aggregations; i.e.,

$$Illiteracy_i^{\text{Census}} = \beta Illiteracy_i^{pop} + X_i'\delta + \varepsilon_i \ , \tag{4}$$

where X_i is a vector of administrative-boundary proxies for illiteracy used as controls (as defined in Footnote 36), and ε_i is the heteroskedasticity-robust error term.

The results in Table 1 show that our newspaper market-based illiteracy proxy is positively and significantly correlated with the actual illiteracy rate across all specifications, regardless of which administrative aggregation proxies are included. In Column (1), we show the binary correlation coefficient. In Column (2), we control for the text-based proxy for illiteracy projected onto municipalities using region boundaries. In Column (3), we augment the specification in Column (2) with the text-based proxy for illiteracy calculated using province boundaries. In Column (4), we add to the specification in Column (3) the proxy for illiteracy calculated using the *circondario* boundaries. Finally, in Column (5), we control for all the text-based proxies projected onto municipalities using administrative aggregations (i.e., we add the text-based proxy calculated used the headquarters' city boundaries). The number of observations drops when adding proxies based on more granular approximations as the latter are computable for fewer municipalities. The estimated coefficients imply that a one standard deviation increase in the interest of

a newspaper for illiteracy is associated with a 0.07–0.2 standard deviation increase in the census illiteracy rate.[37] The correlation between the proxies projected onto municipalities using our methodology and the actual illiteracy rates remains positive and statistically significant when controlling for any proxies based on administrative aggregations. Table A.6 in Appendix A reports virtually unchanged estimates when using surface-weighted illiteracy proxies.

Taken together, these findings indicate that our methodology for estimating historical media markets offers a valuable tool for obtaining proxies of socio-economic variables for which granular data are insufficient or unavailable, or the collection of which would be too costly. The same text-based projection can be applied to any situation in which local-level data are missing, delivering measures with stronger predictive content than administrative-boundary benchmarks.

6 Validation of estimated markets: the *Giro d'Italia*

We test the reliability of our methodology in estimating local newspaper markets by focusing on the coverage of the 1919–1922 editions of the *Giro d'Italia*. The *Giro d'Italia* is one of the most prominent stage races in international professional cycling. The race has taken place yearly in the late spring since 1909 with the only exception of the WWI and WWII years. Although admittedly not a historically crucial event, the features of the race make it the ideal context to assess the reliability of the newspaper markets estimated by our procedure. The *Giro d'Italia* goes through a large number of municipalities – and hence newspaper markets – within a limited period of time. Figure A.14 in Appendix A shows the reconstructed pathways of each edition of the *Giro d'Italia* included in our analysis. The itinerary of the *Giro d'Italia* was relatively stable across the four editions – starting and ending in Milan, where the headquarters of the newspaper organizing the competition (*La Gazzetta dello Sport*, the main sports newspaper in the country still today) is located. The race crossed most metropolitan cities, although many areas of the country were crossed in one edition only.[38] Furthermore, the coverage of the race can be measured by searching for keywords – such as the official name of the competition or the surname of the top cyclists – that do not include any geographic information. These features make the *Giro d'Italia* the ideal setting for estimating a Difference-in-Differences specification that combines the granularity and the high frequency of newspapers' issues.

[37] The specification of Column (5) that controls for the headquarters' text-based illiteracy gives a correlation coefficient of about 0.54. However, this estimate is based on 76 municipalities only and is therefore omitted from the above interval.

[38] See Appendix B for details on the reconstruction of the stage routes for the 1919–1922 editions of *Giro d'Italia*.

Table 2: Newspaper coverage of *Giro d'Italia*

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.:	Giro d'Italia (Binary)			Giro d'Italia (Count)		
Crossed × GiroWeeks	0.186***			1.748***		
	(0.0486)			(0.470)		
CrossedPop × GiroWeeks		0.110***			1.117***	
		(0.0259)			(0.272)	
CrossedSur × GiroWeeks			0.111***			1.051***
			(0.0244)			(0.264)
Observations	17,261	17,261	17,261	17,261	17,261	17,261
R^2	0.298	0.303	0.304	0.266	0.286	0.280
Newspaper-year FE	✓	✓	✓	✓	✓	✓
Week-year FE	✓	✓	✓	✓	✓	✓
Mean dep. var.	0.0648	0.0648	0.0648	0.242	0.242	0.242

Notes: The unit of observation is a newspaper-week. All specifications include newspaper-year fixed effects and week-year fixed effects. Standard errors robust to clustering at the newspaper level are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

We estimate the following regression model:

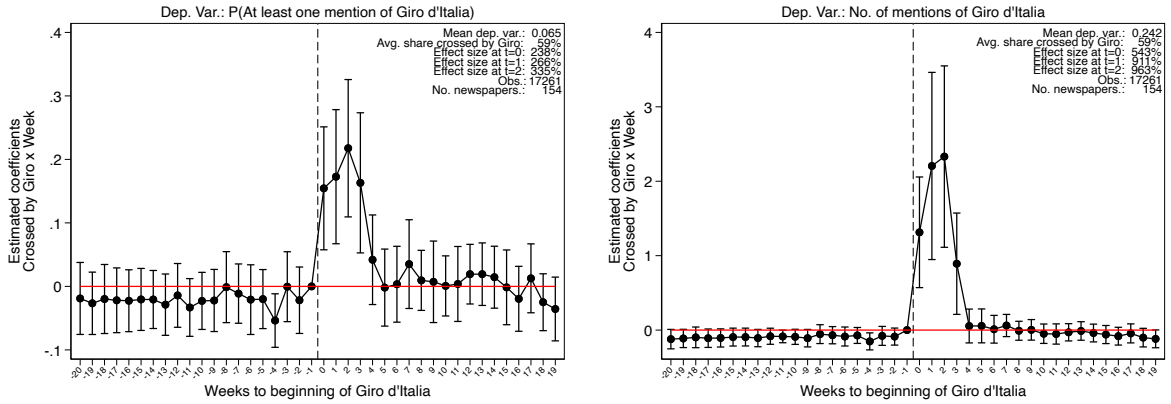
$$Mentions_{j,e,t} = \beta_1 Treated_{j,e} \times GiroWeeks_{e,t} + \psi_{j,e} + \delta_{e,t} + \varepsilon_{j,e,t} \ , \tag{5}$$

where $GiroWeeks_{e,t}$ takes value 1 during the weeks of edition e of the race, $\psi_{j,e}$ is the newspaper-edition (i.e., year) fixed effect, and $\delta_{e,t}$ is the week-edition (i.e., year) fixed effect. With a slight abuse of notation, the independent variable $Treated_{j,e}$ is either the dummy $Crossed_{j,e}$ or the continuous variables $CrossedPop_{j,e}$ and $CrossedSur_{j,e}$. The former takes value 1 if newspaper j is 'crossed' by the race during edition e (i.e., when at least one municipality included in market $m(j)$ is crossed by the *Giro d'Italia* during edition e). The latter represent the shares of $m(j)$'s population and surface, respectively, that are crossed by the race during edition e.[39] The coefficient β_1 identifies the average causal effect of being crossed by the *Giro d'Italia* during the weeks in which the race takes place. We consider as dependent variables both a dummy taking value 1 if newspaper j reports at least one news about the *Giro d'Italia* in the issues published during week t and 0 otherwise, and the total number (count) of mentions of race-related keywords in the issues published by newspaper j during week t. In our main specifications, we compute $Mentions_{j,e,t}$ by searching for the exact name of the race (i.e., *Giro d'Italia*).

We also estimate a dynamic version of Equation (5), in which we replace the indicator

[39] The two variables are calculated as the standardized ratio between the total population (resp., surface) of all municipalities included in $m(j)$ that are crossed by the *Giro d'Italia* and the total population (resp. surface) of $m(j)$.

Figure 9: Newspaper coverage of *Giro d'Italia* – *Dynamic specification*



Notes: The unit of observation is a newspaper-week. All specifications include newspaper-year fixed effects and week-year fixed effects. 95% confidence intervals are based on standard errors robust to clustering at the newspaper level.

variable $GiroWeeks_{e,t}$ with a set of binary indicators for the number of weeks from/to the beginning of edition e of the *Giro d'Italia*. Formally, we estimate the model

$$Mentions_{j,e,t} = \sum_{k=-20, k\neq -1}^{19} \beta_k Treated_{j,e} \times (Weeks\ to\ Giro_{e,t} = k) + \psi_{j,e} + \delta_{e,t} + \varepsilon_{j,e,t}. \tag{6}$$

The estimation of Equation (6) allows us to observe the weekly dynamics of the intensity of the *Giro d'Italia* news coverage, assessing whether there exist parallel trends before the beginning of race, and whether newspapers coverage peaks exactly when the race crosses a given newspaper's market.

The results in Table 2 and Figure 9 strongly corroborate the evidence discussed in Section 5 that our estimation of historical newspaper markets is accurate.[40] We estimate that, exactly at the beginning of the race, there is a sizable spike both in the probability that the newspapers whose market is crossed by the *Giro d'Italia* write about the race, and in their intensity of coverage. This effect quickly disappears following the end of the race.[41]

[40] See Figure A.15 in Appendix A for the results obtained by estimating Equation (6) based on the share of population and the share of surface.

[41] See Table A.7 in Appendix A for the results obtained estimating Equation (5) using as the dependent variable the intensity of mentions of cycling-related keywords or the surnames of top athletes. Table A.8 in Appendix A shows that the share of population or the share of surface crossed by the *Giro d'Italia* during edition e matters only during the race also when focusing only on the intensive margin (i.e., restricting the sample to newspapers the markets of which are crossed by the *Giro d'Italia* during edition e). In Tables A.9 and A.10 in Appendix A, we compare the estimates with those obtained using administrative aggregations to proxy for newspaper markets.

7 Concluding remarks

Historical newspapers contain a wealth of information to measure economic development at a granular and high-frequency resolution. Yet, measuring the geographic market of each newspaper is a challenging task because relevant data (e.g., data on purchases) are usually not available at the local level. This paper outlines a novel data-driven methodology to estimate historical newspaper markets. The standard approach typically assumes that newspaper markets coincide with areas defined by administrative aggregations (such as counties or states). Instead, our technique delivers market areas of endogenous size that do not necessarily overlap with predefined boundaries.

We apply our methodology to data from 154 weekly or daily newspapers published in Italy between 1919 and 1922. Reconstructing historical newspaper markets in those years is challenging because i) the period between the end of WWI and the rise to power of the Fascist movement was exceptionally turbulent; ii) literacy rates were booming throughout the country and many newspapers were founded or ceased operation every year; iii) Italian was not the reference language in some areas of the country.

Our methodology for estimating newspaper markets has two important properties. First, it flexibly determines the number of localities belonging to each market, accounting for the fact that some newspapers may have geographically concentrated or dispersed markets. Second, unlike most of the pertinent literature, it draws the boundaries of newspaper markets independently of administrative aggregations. These properties ensure that the methodology can be applied, regardless of the country or historical period, to obtain markets defined at the lowest possible level of granularity without any prior knowledge about the administrative hierarchy. Our analysis shows that proxies of newspaper markets based on administrative aggregations are often inaccurate, systematically underestimating or overestimating actual markets. Instead, the text-based measures of economic development projected at the local level using the markets identified by our procedure turn out to be stronger predictors of actual economic variables than those obtained by projecting onto the territory the same text-based measures based on alternative approaches.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High Wage Workers and High Wage Firms. *Econometrica 67*(2), 251–333.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the Rise of the Nazis in Prewar Germany. *The Quarterly Journal of Economics 130*(4), 1885–1939.

Banerjee, A., E. Duflo, and N. Qian (2020). On the Road: Access to Transportation Infrastructure and Economic Growth in China. *Journal of Development Economics 145*, 102442.

Beach, B. and W. W. Hanlon (2023a). Culture and the Historical Fertility Transition. *The Review of Economic Studies 90*(4), 1669–1700.

Beach, B. and W. W. Hanlon (2023b). Historical Newspaper Data: A Researcher's Guide. *Explorations in Economic History 90*, 101541.

Bhuller, M., T. Havnes, J. McCauley, and M. Mogstad (2024, April). How the Internet Changed the Market for Print Media. *American Economic Journal: Applied Economics 16*(2), 318–58.

Cagé, J. (2020). Media Competition, Information Provision and Political Participation: Evidence from French Local Newspapers and Elections, 1944–2014. *Journal of Public Economics 185*, 104077.

Calderon, A., V. Fouka, and M. Tabellini (2023). Racial Diversity and Racial Policy Preferences: the Great Migration and Civil Rights. *The Review of Economic Studies 90*(1), 165–200.

CENSIS (2004). Cenni di Storia dei Sistemi di Informazione e Comunicazione.

Combes, P.-P., L. Gobillon, and Y. Zylberberg (2022). Urban Economics in a Historical Perspective: Recovering Data with Machine Learning. *Regional Science and Urban Economics 94*, 103711.

Correia, S. and S. Luck (2023). Digitizing Historical Balance Sheet Data: A Practitioner's Guide. *Explorations in Economic History 87*, 101475.

Dehdari, S. H. and K. Gehring (2022). The Origins of Common Identity: Evidence from Alsace-Lorraine. *American Economic Journal: Applied Economics 14*(1), 261–292.

Dell, M. (2010). The Persistent Effects of Peru's Mining Mita. *Econometrica 78*(6), 1863–1903.

Dell, M. (2025). Deep Learning for Economists. *Journal of Economic Literature 63*(1), 5–58.

Dell, M., J. Carlson, T. Bryan, E. Silcock, A. Arora, Z. Shen, L. D'Amico-Wong, Q. Le, P. Querubin, and L. Heldring (2023). American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers.

DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014). Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia. *American Economic Journal: Applied Economics 6*(3), 103–132.

Djourelova, M., R. Durante, and G. J. Martin (2024). The Impact of Online Competition on Local Newspapers: Evidence from the Introduction of Craigslist. *The Review of Economic Studies 92*(3), 1738–1772.

Doucette, J. (2024). Parliamentary Constraints and Long-Term Development: Evidence from the Duchy of Württemberg. *American Journal of Political Science 68*(1), 24–41.

Drago, F., T. Nannicini, and F. Sobbrio (2014). Meet the press: How Voters and Politicians Respond to Newspaper Entry and Exit. *American Economic Journal: Applied Economics 6*(3), 159–188.

Durante, R., P. Pinotti, and A. Tesei (2019). The Political Legacy of Entertainment TV. *American Economic Review 109*(7), 2497–2530.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011). Media and Political Persuasion: Evidence from Russia. *American Economic Review 101*(7), 3253–85.

Faber, B. (2014). Trade Integration, Market size, and Industrialization: Evidence from China's National Trunk Highway System. *The Review of Economic Studies 81*(3), 1046–1070.

Fan, Y. (2013). Ownership Consolidation and Product Characteristics: A Study of the US Daily Newspaper Market. *The American Economic Review 103*(5), 1598–1628.

Feigenbaum, J. and D. P. Gross (2024). Answering the Call of Automation: How the Labor Market Adjusted to Mechanizing Telephone Operation. *The Quarterly Journal of Economics 139*(3), 1879–1939.

Ferguson-Cradler, G. (2023). Narrative and Computational Text Analysis in Business and Economic History. *Scandinavian Economic History Review 71*(2), 103–127.

Ferrara, A., J. Y. Ha, and R. Walsh (2024). Using Digitized Newspapers to Address Measurement Error in Historical Data. *The Journal of Economic History 84*(1), 271–306.

Fontana, N., T. Nannicini, and G. Tabellini (2023). Historical roots of political extremism: The effects of Nazi occupation of Italy. *Journal of Comparative Economics 51*(3), 723–743.

Galvis, Á. F., J. M. Snyder Jr, and B. Song (2016). Newspaper Market Structure and Behavior: Partisan Coverage of Political Scandals in the United States from 1870 to 1910. *The Journal of Politics 78*(2), 368–381.

Gentzkow, M. and J. M. Shapiro (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011). The Effect of Newspaper Entry and Exit on Electoral Politics. *American Economic Review 101*(7), 2980–3018.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2014). Competition and Ideological Diversity: Historical Evidence from US Newspapers. *American Economic Review 104*(10), 3073–3114.

George, L. M. and J. Waldfogel (2006). The New York Times and the Market for Local Newspapers. *American Economic Review 96*(1), 435–447.

Hirano, S. and J. M. Snyder (2024). Measuring the Partisan Behavior of U.S. Newspapers, 1880 to 1980. *The Journal of Economic History 84*(2), 554–592.

Ministero di Agricoltura, Industria e Commercio (1908). *Annuario Statistico Italiano 1905–1907*. Roma: Tipografia Nazionale di G. Bertero e C.

Olken, B. A. (2009). Do Television and Radio Destroy Social Capital? Evidence from Indonesian Villages. *American Economic Journal: Applied Economics 1*(4), 1–33.

Ottinger, S. and M. Posch (2022). The Political Economy of Propaganda: Evidence from U.S. Newspapers. IZA DP No. 15078.

Perlman, E. R. and S. Sprick Schuster (2016). Delivering the Vote: The Political Effect of Free Mail Delivery in Early Twentieth Century America. *The Journal of Economic History 76*(3), 769–802.

Petrova, M. (2011). Newspapers and Parties: How Advertising Revenues Created an Independent Press. *The American Political Science Review 105*(4), 790–808.

Puglisi, R. and J. M. J. Snyder (2011). Newspaper Coverage of Political Scandals. *The Journal of Politics 73*(3), 931–950.

Riley, S. J., S. D. DeGloria, and R. Elliot (1999). Index that Quantifies Topographic Heterogeneity. *Intermountain Journal of Sciences 5*(1-4), 23–27.

Sardoschau, S., G. Gulino, and F. Masera (2025). Identity Under Scrutiny: Media Attention and Rule Compliance. IZA DP No. 17888.

Scheve, K. and T. Serlin (2023). The German Trade Shock and the Rise of the Neo-Welfare State in Early Twentieth-Century Britain. *American Political Science Review 117*(2), 557–574.

Scheve, K. and T. Serlin (2025). Trains, Trade, and Transformation: A Spatial Rogowski Theory of America's 19th-century Protectionism. *American Journal of Political Science 69*(3), 915–929.

Seamans, R. and F. Zhu (2014). Responses to Entry in Multi-sided Markets: The Impact of Craigslist on Local Newspapers. *Management Science 60*(2), 476–493.

Shen, Z., R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, and W. Li (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. arXiv preprint arXiv:2103.15348.

Snyder, J. M. J. and D. Strömberg (2010). Press Coverage and Political Accountability. *Journal of Political Economy 118*(2), 355–408.

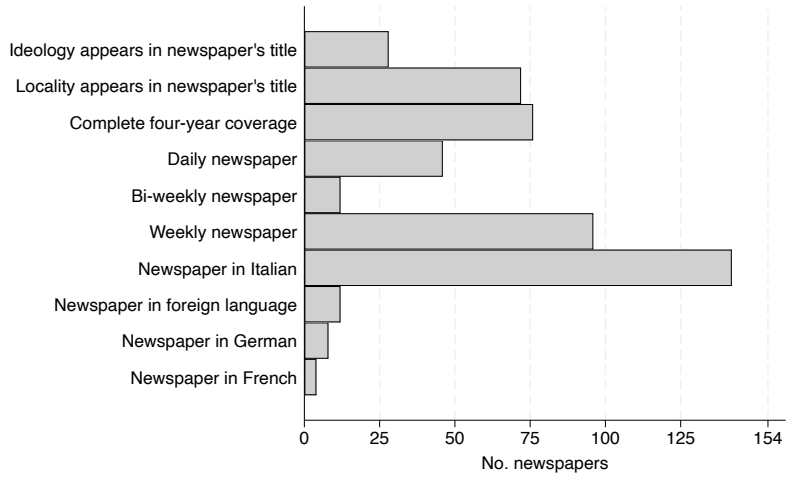
Touring Club Italiano (1913). *Carta d'Italia del Touring club italiano*. Novara: Istituto Geografico De Agostini.

Appendix

A Figures and Tables

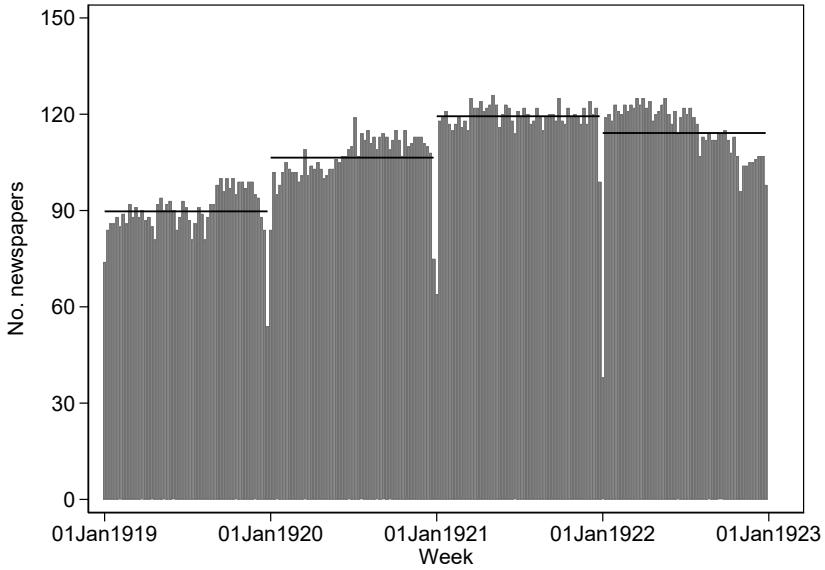
Figures

Figure A.1: Characteristics of newspapers in the sample



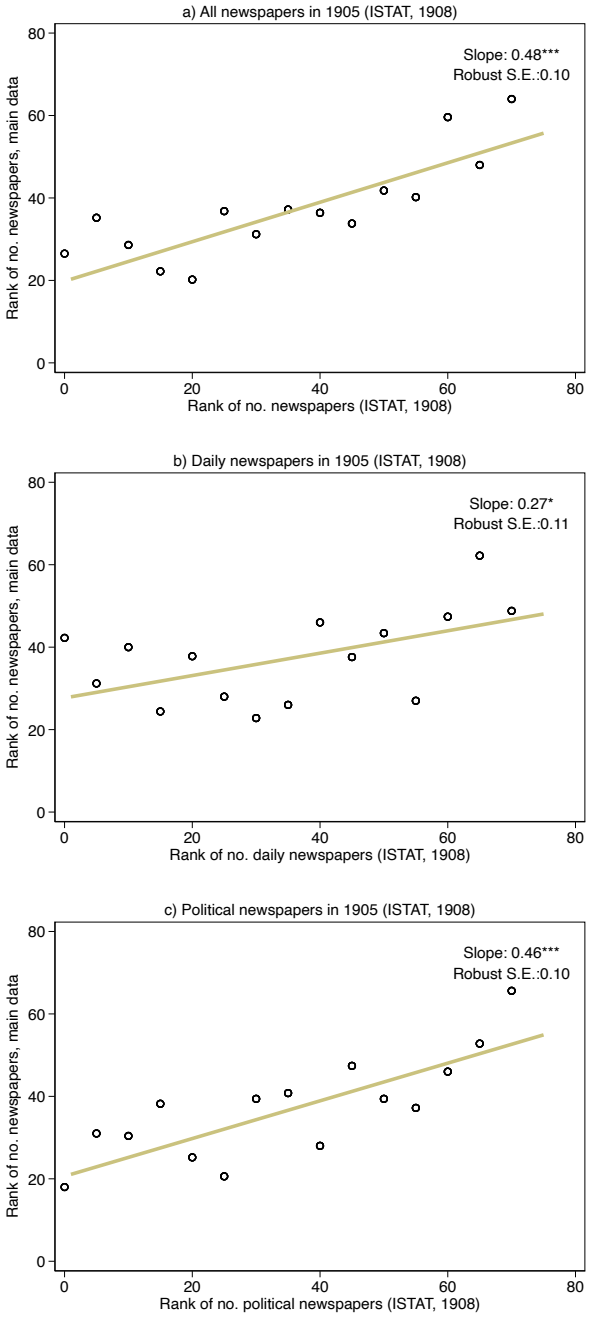
Notes: The unit of observation is a newspaper. Source for this information: authors' manual inspection of digitized newspapers in the sample.

Figure A.2: Number of available newspapers per week



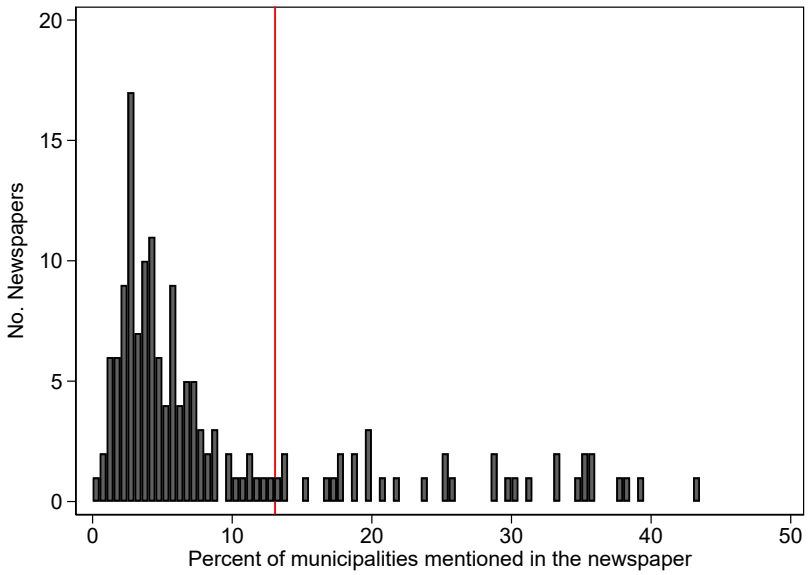
Notes: The unit of observation is a newspaper. Each bar reports the number of newspapers for which we observe at least one issue published during the week specified on the horizontal axis. Horizontal black lines represent the average number of newspapers that we observe in each year.

Figure A.3: Correlation between headquarters of available newspapers and headquarters of newspapers registered in Ministero di Agricoltura, Industria e Commercio (1908)



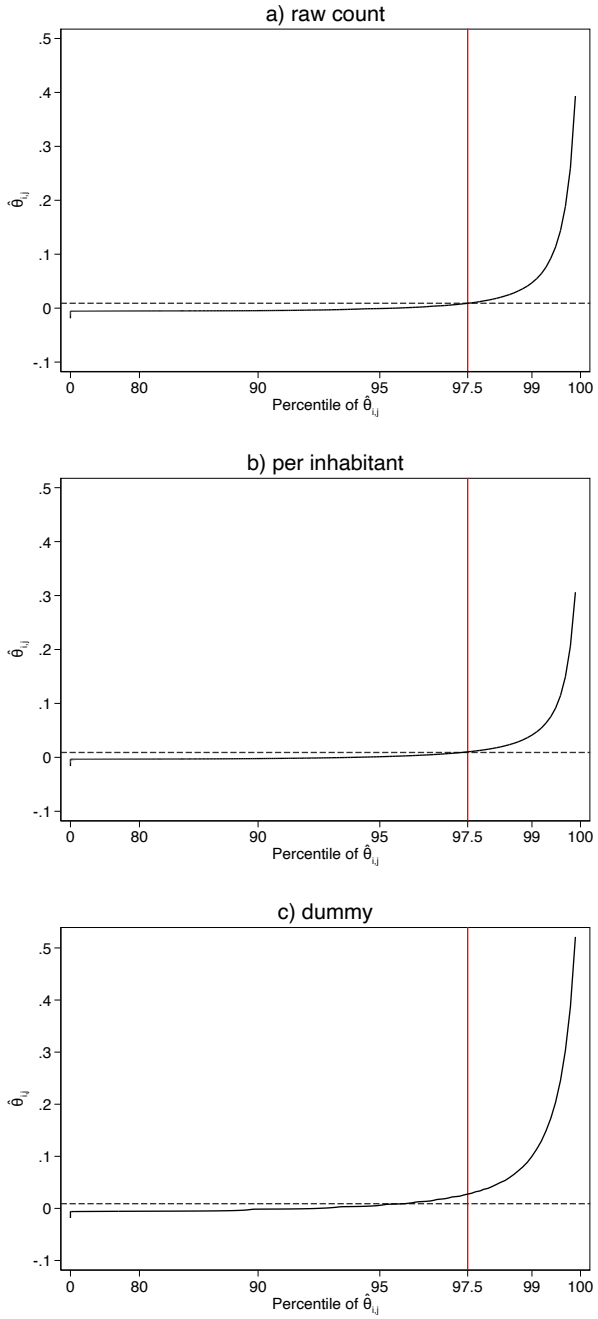
Notes: The unit of observation is a province (1921 boundaries). Panel (a) reports the binary correlation between the (rank of the) number of newspaper headquarters per province in our sample and the (rank of the) number of newspaper headquarters per province according to Ministero di Agricoltura, Industria e Commercio (1908). Panel (b) reports the binary correlation between the (rank of the) number of newspaper headquarters per province in our sample and the (rank of the) number of daily newspaper headquarters per province according to Ministero di Agricoltura, Industria e Commercio (1908). Panel (c) reports the binary correlation between the (rank of the) number of newspaper headquarters per province in our sample and the (rank of the) number of political newspaper headquarters per province according to Ministero di Agricoltura, Industria e Commercio (1908). In all panels, markers represent sample averages of the (rank of the) number of newspaper headquarters per province in our sample within bins of five provinces, ranked according to the measure specified in the horizontal axis. Standard errors are robust to heteroskedasticity. N=75.

Figure A.4: Number of municipalities mentioned by each newspaper



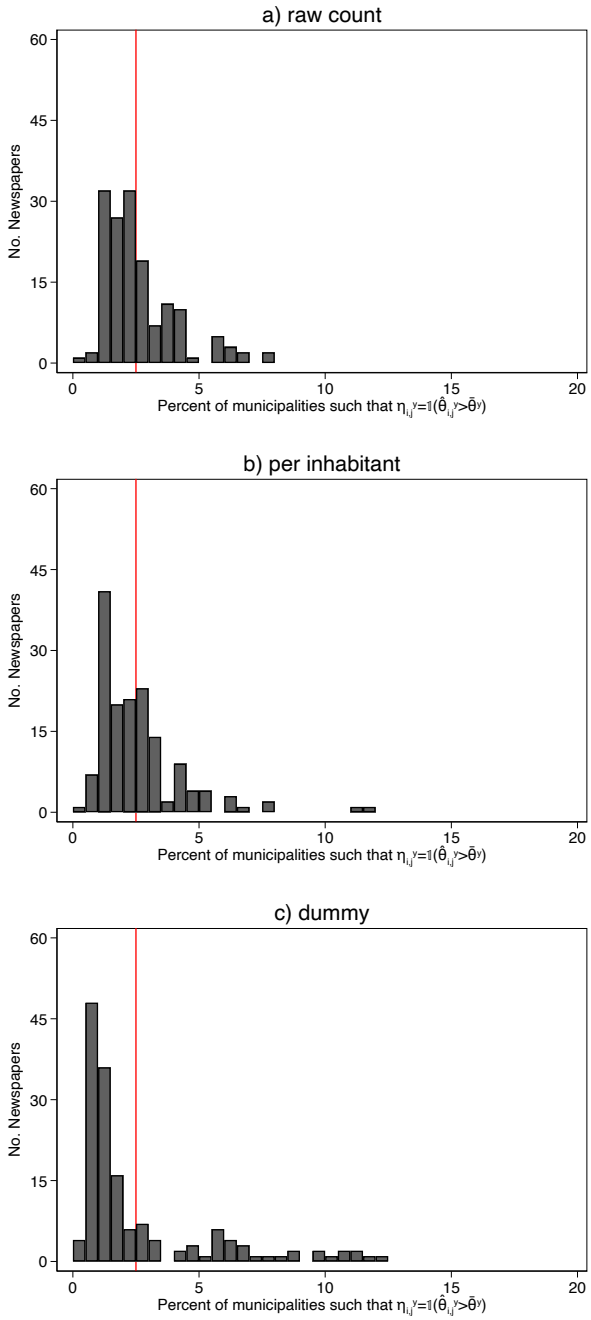
Notes: The unit of observation is a newspaper. This figure reports the distribution of the number of municipalities mentioned at least once by each newspaper. The vertical red line represents the average number of mentions.

Figure A.5: Estimation output – Empirical distributions Θ^y



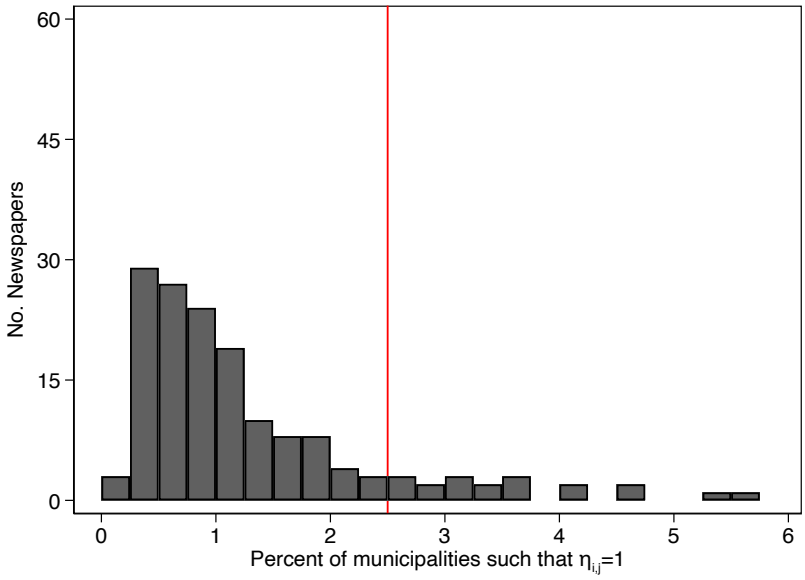
Notes: The unit of observation is a municipality-newspaper tuple. N=1,219,834. The figure reports the minimum value of $\widehat{\theta^y_{i,j}}$ within bins sized 0.1 percent of the distribution Θ^y for each dependent variable. In Panel (a), the dependent variable is the raw number of mentions; in Panel (b), the dependent variable is the number of mentions per inhabitant; in Panel (c), the dependent variable is a dummy for at least one mention. To ease the visualization, the bottom 0.1 percent and the top 0.1 of the distribution are not reported, and the horizontal axis is reported on a non-linear scale. The vertical red line represents the 97.5 percentile, while the horizontal dashed line reports its corresponding value (i.e., the value of the threshold $\overline{\theta}^y$ in Condition (2)).

Figure A.6: Estimation output – Distribution of the share of municipality-newspapers tuples such that $\eta_{i,j}^{y} = 1$



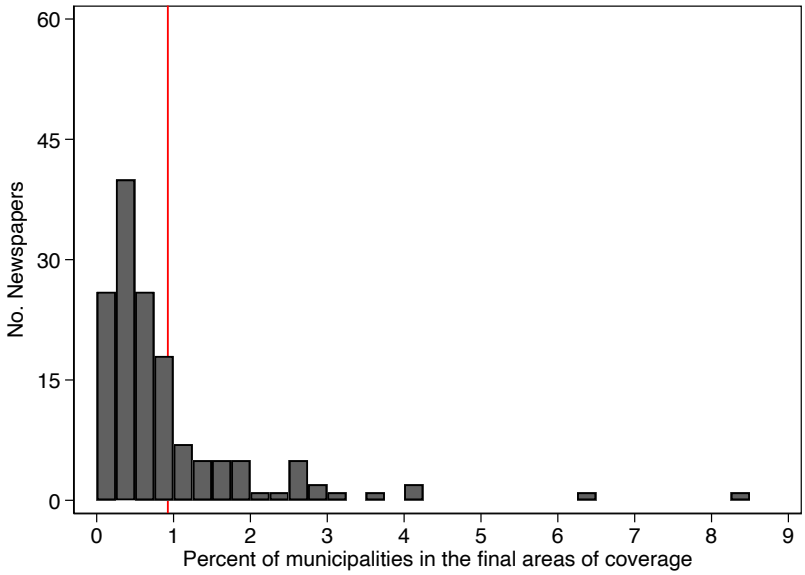
Notes: The unit of observation is a municipality-newspaper tuple. N=1,219,834. The figure reports the distribution of the number of municipalities such that $\eta_{i,j}^{y} = 1$ for each newspaper j. The vertical red line represents the 97.5 percentile.

Figure A.7: Estimation output – Distribution of the share of municipality-newspapers tuples such that $\eta_{i,j} = 1$



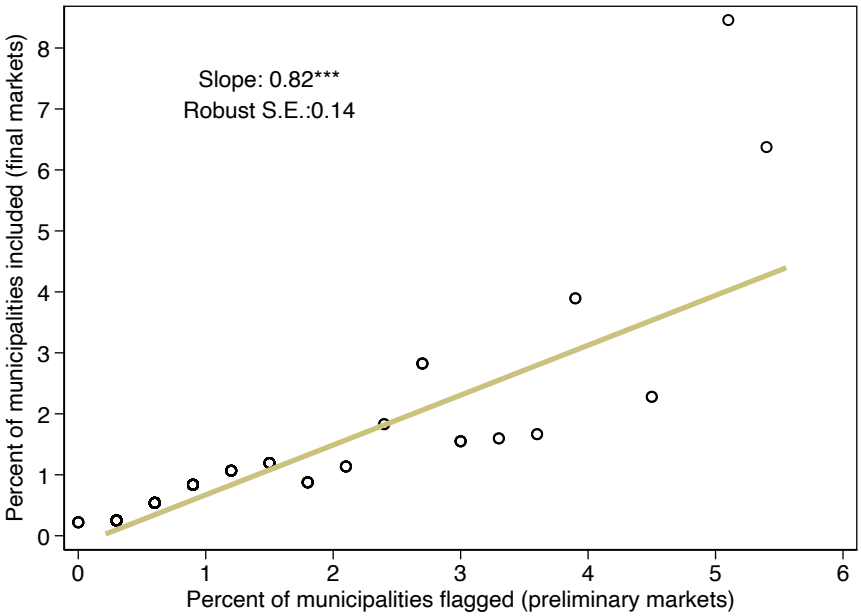
Notes: The unit of observation is a municipality-newspaper tuple. N=1,219,834. The figure reports the distribution of the number of municipalities such that $\eta_{i,j} = 1$ for each newspaper j. The vertical red line represents the average preliminary market size.

Figure A.8: Distribution of estimated newspaper markets' size



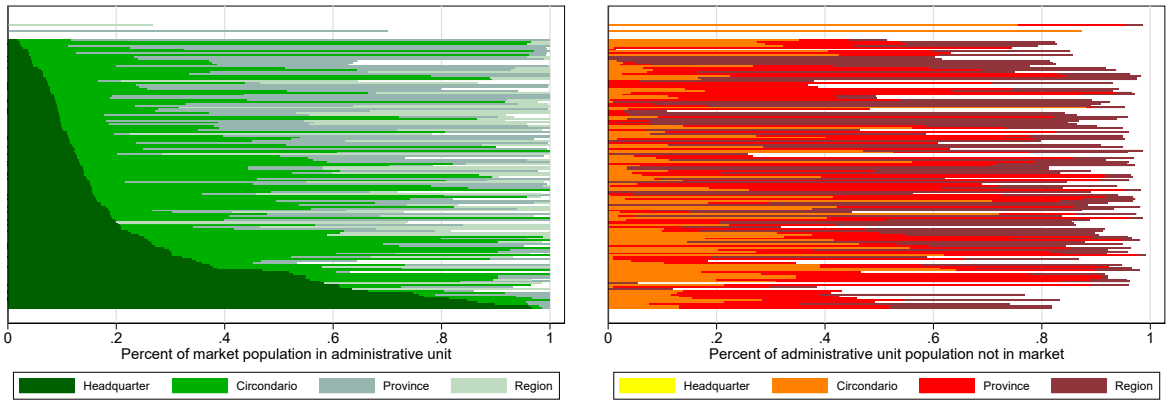
Notes: The unit of observation is a municipality-newspaper tuple. The figure reports the distribution of the number of municipalities belonging to each estimated newspaper market. The vertical red line represents the average market size (0.93% of municipalities).

Figure A.9: Correlation between the size of preliminary markets and the size of final markets



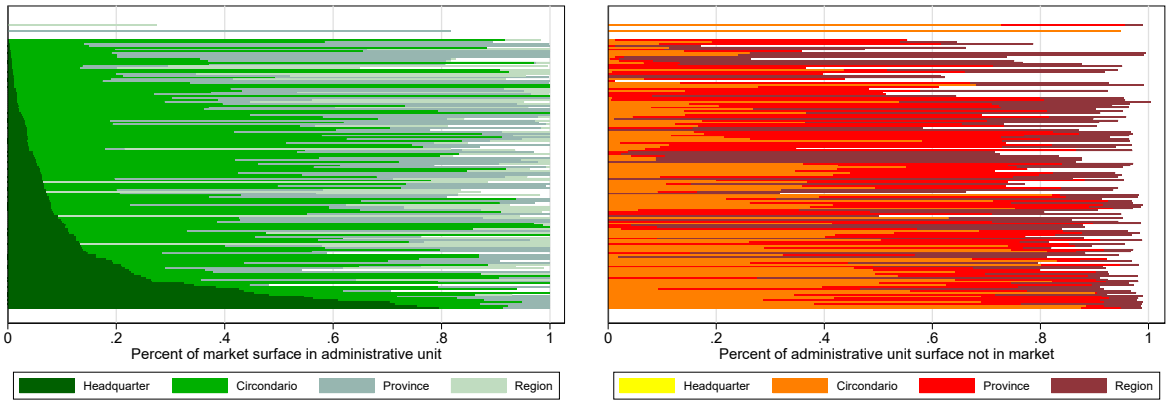
Notes: The unit of observation is a newspaper. The graph shows the binary correlation between the share of municipalities included in each estimated market and the share of municipalities flagged for inclusion in the preliminary market. Markers represent sample averages of the share of municipalities in each final market within bins of 0.3 percentage points in the preliminary share. N=147.

Figure A.10: Population discrepancy between estimated markets and proxies based on administrative aggregations



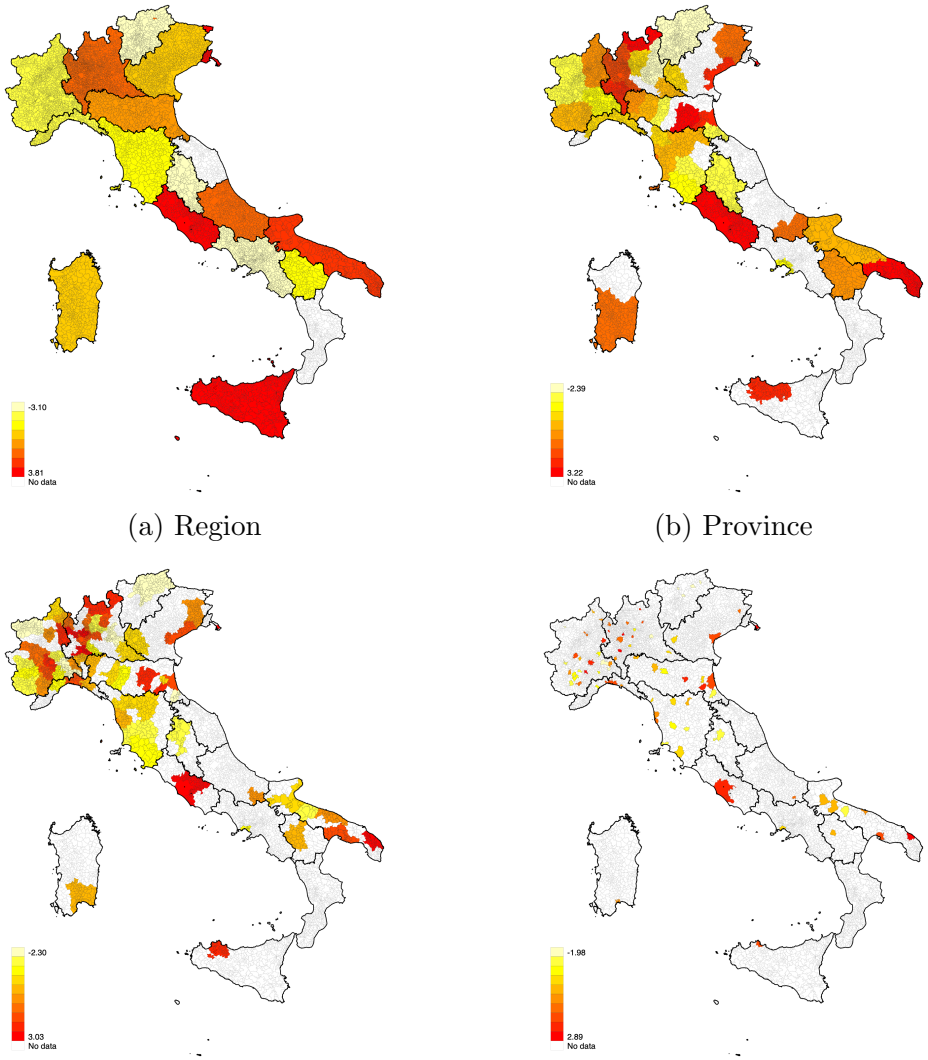
Notes: The unit of observation is a newspaper. Panel (a) reports the cumulative percent of estimated market population which belongs to the headquarters' city, the *circondario* of the headquarters' city, the province of the headquarters' city, and the region of the headquarters' city, respectively. Panel (b) reports the percentage of administrative unit population (among headquarters' city, *circondario* of the headquarters, province of the headquarters, and region of the headquarters) which does not belong to estimated newspaper market.

Figure A.11: Surface discrepancy between estimated markets and proxies based on administrative aggregations



Notes: The unit of observation is a newspaper. Panel (a) reports the cumulative percent of estimated market surface which belongs to the headquarters' city, the *circondario* of the headquarters' city, the province of the headquarters' city, and the region of the headquarters' city, respectively. Panel (b) reports the percentage of administrative unit surface (among headquarters' city, *circondario* of the headquarters, province of the headquarters, and region of the headquarters) which does not belong to estimated newspaper market.

Figure A.12: Illiteracy rate – Projection of textual extraction based on administrative aggregations

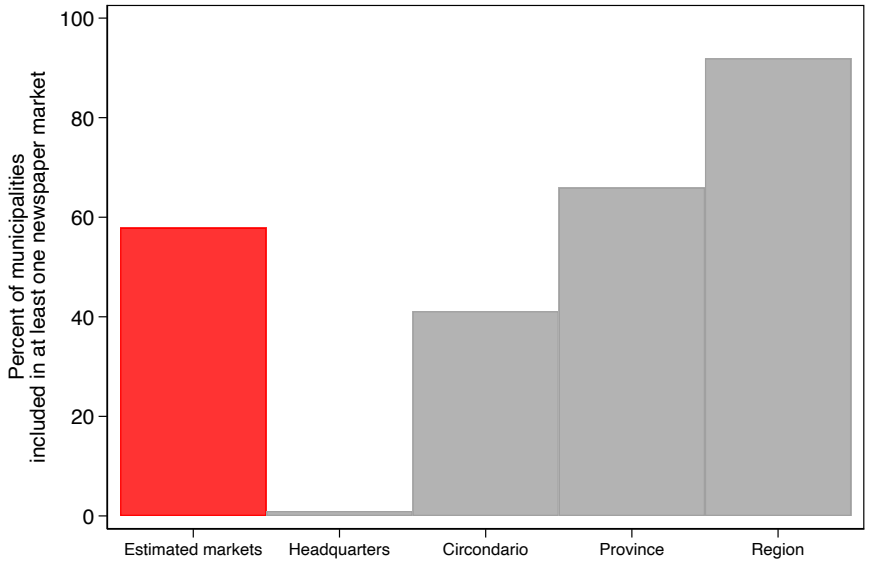


(a) Region (b) Province

(c) Circondario (d) Headquarters

Notes: In all panels, illiteracy measures are standardized to have mean 0 and standard deviation 1. Each color represents one decile in the empirical distribution, with darker colors capturing higher intensity of illiteracy-related keywords in the newspapers' texts. Panel (a) plots the standardized incidence of illiteracy-related keywords projected onto municipalities using 1921 region boundaries. Panel (b) plots the standardized incidence of illiteracy-related keywords projected onto municipalities using 1921 province boundaries. Panel (c) plots the standardized incidence of illiteracy-related keywords projected onto municipalities using 1921 *circondario* boundaries. Panel (d) plots the standardized incidence of illiteracy-related keywords projected onto municipalities using headquarters' city boundaries.

Figure A.13: Number of municipalities included in at least one newspaper market



Notes: The unit of observation is a municipality (1991 boundaries, see Footnote 26 for details). Each bar reports the percentage of municipalities that result as being a part of at least one newspaper's market, based on our estimates (red bar) or administrative aggregations (gray bars).

Figure A.14: *Giro d'Italia* reconstructed stage routes 1919–1922

(a) 1919

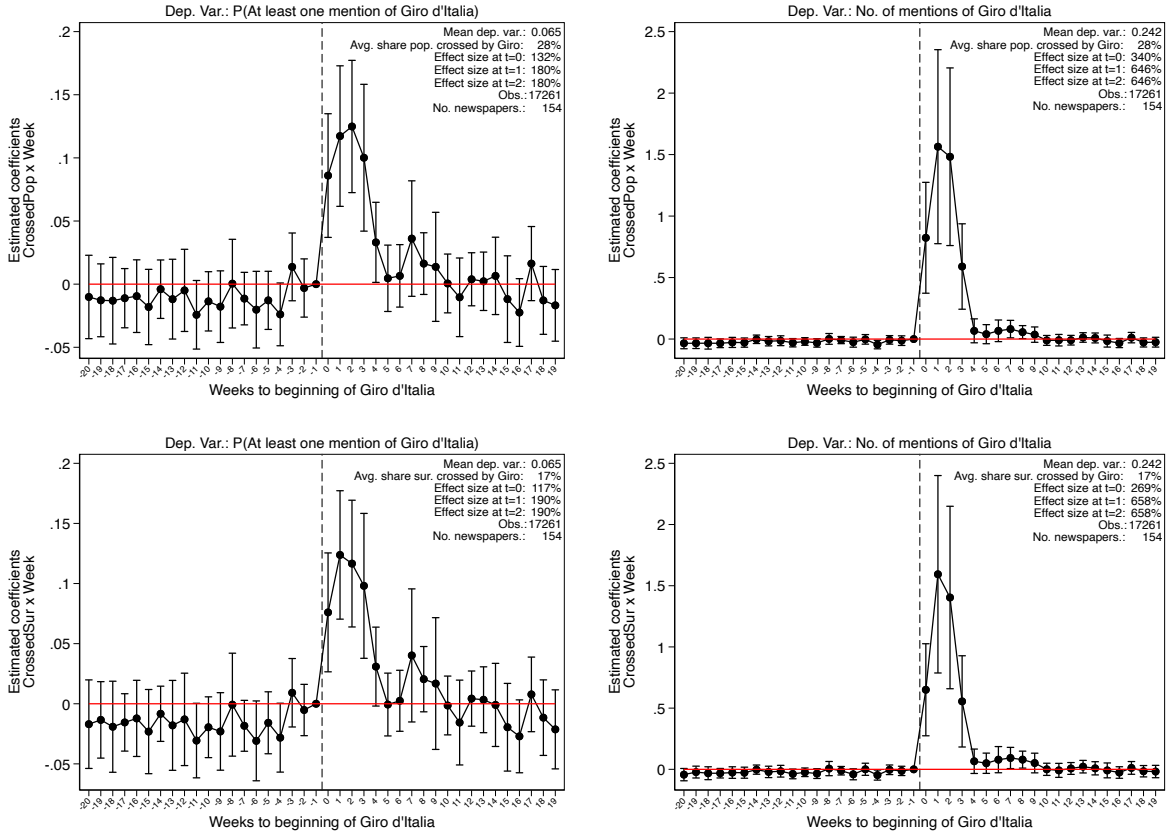
(b) 1920

(c) 1921

(d) 1922

Notes: The maps illustrate the reconstructed stage routes of the *Giro d'Italia* in the years 1919 (Panel (a)), 1920 (Panel (b)), 1921 (Panel (c)), and 1922 (Panel (d)). Each red dot reports a municipality that is explicitly mentioned in the daily reports published by the *La Gazzetta dello Sport*. The black lines represent the predicted routes of each stage, constructed by minimizing the travel costs between two consecutive localities (see Appendix B for details). The territory that corresponds to contemporary Slovenia and Switzerland is included because some stages crossed the Istria peninsula (part of the Kingdom of Italy in the interwar years) and some localities within the Italian-speaking Swiss canton of Ticino.

Figure A.15: Newspaper coverage of *Giro d'Italia – Dynamic specification*



Notes: The unit of observation is a newspaper-week. All specifications include newspaper-year fixed effects and week-year fixed effects. 95% confidence intervals are based on standard errors robust to clustering at the newspaper level.

Tables

Table A.1: Determinants of newspaper headquarters

	(1)	(2)	(3)	(4)
Dep. var.:	1 = Newspaper headquarter		No. newspapers' headquarters	
log(Population), std.	0.0287***	0.0358***	0.0628***	0.0771***
	(0.00334)	(0.00367)	(0.00977)	(0.0102)
log(Surface), std.	–0.00415*	–0.00690***	–0.00783	–0.0126**
	(0.00209)	(0.00251)	(0.00494)	(0.00552)
% Literate, std.	0.0101***	0.0106***	0.0210***	0.0260***
	(0.00267)	(0.00287)	(0.00627)	(0.00894)
% Entrepreneurs, std.	0.00565***	0.00203	0.0162***	0.00979***
	(0.00152)	(0.00124)	(0.00377)	(0.00362)
% Industrial workers, std.	–0.00109	0.00194	–0.00177	0.00385
	(0.00224)	(0.00190)	(0.00473)	(0.00466)
% Unemployed, std.	–0.000173	–0.00167**	–0.00159	–0.00471***
	(0.00102)	(0.000814)	(0.00200)	(0.00169)
Observations	9,195	9,195	9,195	9,195
R^2	0.082	0.098	0.067	0.078
Region FE		✓		✓
Mean dep. var.	0.00837	0.00837	0.0167	0.0167

Notes: The unit of observation is a municipality (1921 boundaries). All regressors are measured at the 1921 Census and are standardized to have mean equal to 0 and variance equal to 1. Standard errors robust to clustering at the province level (1921 boundaries) are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

Table A.2: Descriptive statistics of the raw frequency of mentions

	mean	sd	min	max
Raw count	.0072937	.2299131	0	131
Per inh.	9.07e-07	.0000404	0	.1034483
Dummy	.0032847	.0572185	0	1
Observations	495965494			

Notes: The unit of observation is a municipality-newspaper-week. This table reports the descriptive statistics of the three measures of the raw frequency of mentions which we use as dependent variable to estimate Equation (1).

Table A.3: Correlation between the three raw measures of newspaper coverage of each municipality

	Raw count	Per inh.	Dummy
Raw count	1		
Per inh.	0.230	1	
Dummy	0.553	0.391	1
Observations	495965494		

Notes: The unit of observation is a municipality-newspaper-week. This table reports the binary correlation coefficients between each of the three measures of the raw frequency of mentions which we use as dependent variable to estimate Equation (1).

Table A.4: Descriptive statistics of the empirical distribution Θ

Statistic	raw count	per inhabitant	dummy
Observations	1219834	1219834	1219834
Mean	-.0003184	-6.96e-08	-.0007448
St. Dev.	.1493942	.0000129	.042853
Skewness	87.30867	65.95338	14.90031
Kurtosis	10223.72	8002.351	269.1307
Minimum	-1.415844	-.0001025	-.1410173
Maximum	28.11747	.002726	1.036001
p(25)	-.0058856	-8.59e-07	-.0061543
p(50)	-.0058162	-8.42e-07	-.0060472
p(75)	-.0053379	-7.67e-07	-.0056166
p(90)	-.0045827	-5.12e-07	-.0014059
p(95)	-.0005993	3.10e-07	.0066803
p(97.5)	.009036	2.34e-06	.0274675
p(99)	.0544895	.0000108	.1133958
p(99.5)	.1135374	.000021	.2038784

Notes: The unit of observation is a municipality-newspaper. This table reports the descriptive statistics of the empirical distribution Θ obtained estimating Equation (1).

Table A.5: Correlation between the three $\eta^{y}_{i,j}$

	Raw count	Per inh.	Dummy
Raw count	1		
Per inh.	0.611	1	
Dummy	0.742	0.557	1
Observations	1219834		

Notes: The unit of observation is a municipality-newspaper. This table reports the binary correlation coefficients between each of the three $\eta^{y}_{i,j}$ obtained from the estimation of Equation (1).

Table A.6: Correlation between actual illiteracy rate and textual extraction (surface)

	(1)	(2)	(3)	(4)	(5)
Dep. var.:	Illiteracy rate (1921 census)				
Illit. newsp. incid.	0.208***	0.0790***	0.113***	0.178***	0.526***
	(0.0119)	(0.0155)	(0.0195)	(0.0293)	(0.144)
Observations	4,594	4,515	3,970	2,876	76
R^2	0.052	0.095	0.105	0.120	0.249
Region text-based illiteracy		✓	✓	✓	✓
Province text-based illiteracy			✓	✓	✓
Circondario text-based illiteracy				✓	✓
Headquarters text-based illiteracy					✓
Mean dep. var.	–0.243	–0.255	–0.312	–0.492	–0.553

Notes: The unit of observation is a municipality (1991 boundaries, see Footnote 26 for details). Column (1) reports the correlation between the standardized illiteracy rate (1921 census) and the standardized incidence of illiteracy-related keywords, projected onto municipalities using estimated markets. Columns (2)–(5) augment the specification with controls for the standardized incidence of illiteracy-related keywords projected onto municipalities using region boundaries, province boundaries, *circondario* boundaries, and headquarters' city boundaries, respectively. Standard errors robust to heteroskedasticity are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

Table A.7: Alternative measures of newspaper coverage of *Giro d'Italia*

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. var.:	Cycling/Cyclists		Riders' surnames		All keywords	
	Binary	Count	Binary	Count	Binary	Count
Crossed × GiroWeeks	0.0374**	0.255***	0.0443*	9.236***	0.0499***	11.24***
	(0.0179)	(0.0816)	(0.0236)	(2.576)	(0.0189)	(3.076)
Observations	17,261	17,261	17,261	17,261	17,261	17,261
R^2	0.451	0.434	0.520	0.337	0.562	0.370
Mean dep. var.	0.248	0.565	0.296	2.080	0.402	2.887
Newspaper-year FE	✓	✓	✓	✓	✓	✓
Week-year FE	✓	✓	✓	✓	✓	✓

Notes: The unit of observation is a newspaper-week. All specifications include newspaper-year fixed effects and week-year fixed effects. Standard errors robust to clustering at the newspaper level are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

Table A.8: Newspaper coverage of *Giro d'Italia* – Markets crossed by the *Giro*

	(1)	(2)	(3)	(4)
Dep. var.:	Giro d'Italia (Binary)		Giro d'Italia (Count)	
CrossedPop × GiroWeeks	0.0937**		1.078**	
	(0.0437)		(0.424)	
CrossedSur × GiroWeeks		0.0893**		0.859**
		(0.0354)		(0.427)
Observations	10,116	10,116	10,116	10,116
R^2	0.327	0.329	0.315	0.311
Newspaper-year FE	✓	✓	✓	✓
Week-year FE	✓	✓	✓	✓
Mean dep. var.	0.0830	0.0830	0.336	0.336

Notes: The unit of observation is a newspaper-week. All specifications include newspaper-year fixed effects and week-year fixed effects. Standard errors robust to clustering at the newspaper level are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

Table A.9: Newspaper coverage of *Giro d'Italia* – Comparison with proxies based on administrative aggregations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var.:	Giro d'Italia (Binary)							
CrossedPop × GiroWeeks	0.0325	–0.0172	0.114***	0.105***				
	(0.0341)	(0.0363)	(0.0379)	(0.0314)				
std(Share headquarters pop. crossed) × Giro	0.119***							
	(0.0303)							
std(Share circondario pop. crossed) × Giro		0.167***						
		(0.0358)						
std(Share province pop. crossed) × Giro			–0.00593					
			(0.0363)					
std(Share region pop. crossed) × Giro				0.00833				
				(0.0301)				
CrossedSur × GiroWeeks					0.0404	0.0438	0.133***	0.103***
					(0.0336)	(0.0365)	(0.0309)	(0.0283)
std(Share headquarters sur. crossed) × Giro					0.116***			
					(0.0301)			
std(Share circondario sur. crossed) × Giro						0.0949***		
						(0.0318)		
std(Share province sur. crossed) × Giro							–0.0346	
							(0.0309)	
std(Share region sur. crossed) × Giro								0.0148
								(0.0286)
Observations	17,261	17,261	17,261	17,261	17,261	17,261	17,261	17,261
R^2	0.316	0.321	0.303	0.304	0.316	0.311	0.305	0.304
Newspaper-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Week-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Mean dep. var.	0.0648	0.0648	0.0648	0.0648	0.0648	0.0648	0.0648	0.0648

Notes: The unit of observation is a newspaper-week. All specifications include newspaper-year fixed effects and week-year fixed effects. Standard errors robust to clustering at the newspaper level are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

Table A.10: Newspaper coverage of *Giro d'Italia* – Comparison with proxies based on administrative aggregations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. var.:	Giro d'Italia (Count)							
CrossedPop × GiroWeeks	0.552*	–0.102	0.850***	0.996***				
	(0.305)	(0.336)	(0.312)	(0.326)				
std(Share headquarters pop. crossed) × Giro	0.870***							
	(0.327)							
std(Share circondario pop. crossed) × Giro		1.599***						
		(0.434)						
std(Share province pop. crossed) × Giro			0.380					
			(0.281)					
std(Share region pop. crossed) × Giro				0.215				
				(0.288)				
CrossedSur × GiroWeeks					0.479	0.335	1.217***	0.852***
					(0.404)	(0.333)	(0.337)	(0.301)
std(Share headquarters sur. crossed) × Giro					0.934**			
					(0.416)			
std(Share circondario sur. crossed) × Giro						1.009***		
						(0.288)		
std(Share province sur. crossed) × Giro							–0.264	
							(0.271)	
std(Share region sur. crossed) × Giro								0.356
								(0.292)
Observations	17,261	17,261	17,261	17,261	17,261	17,261	17,261	17,261
R^2	0.302	0.327	0.288	0.287	0.301	0.300	0.282	0.284
Newspaper-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Week-year FE	✓	✓	✓	✓	✓	✓	✓	✓
Mean dep. var.	0.242	0.242	0.242	0.242	0.242	0.242	0.242	0.242

Notes: The unit of observation is a newspaper-week. All specifications include newspaper-year fixed effects and week-year fixed effects. Standard errors robust to clustering at the newspaper level are reported in parentheses. Labels *, **, and *** denote significance levels of 10%, 5%, and 1%, respectively.

B Reconstruction of the yearly itinerary of the *Giro d'Italia*

We reconstruct each year's itinerary of the *Giro d'Italia* accessing the official reports published historically by the *La Gazzetta dello Sport*, which is both the main sports newspaper published in Italy and the organizer of the cycling race. Each daily report on the race published by the *La Gazzetta dello Sport* indicates all the main localities crossed on that specific day (*tappa*). For each race day, several localities (typically between 10–30 kilometers from each other) are cited in the daily report. As is common in the literature on infrastructure economics, we predict the road connecting each pair of consecutive localities by calculating the least costly connection segment taking into account the characteristics of the terrain (see, e.g., Banerjee et al., 2020; Faber, 2014).

More specifically, we construct a cost surface to simulate stage routes based on terrain ruggedness and historical road networks. First, we use a digital elevation model (DEM) raster covering the European continent, with elevation data at a resolution of 90 m×90 m.[42] We pre-process the raster by removing all cells corresponding to large water bodies – including lakes, rivers, and lagoons – to ensure that the simulated paths do not cross natural obstacles. We then use the QGIS Ruggedness tool to calculate a cell-level terrain ruggedness index (TRI) following Riley et al. (1999). To reduce the variance in the TRI values and prevent the simulated paths from being overly sensitive to minor elevation changes, we discretize the values into 30 uniform classes, assigning a cost value between 1 and 30 (with lower values representing easier terrain) to each class. The final cost raster is obtained by overlaying this grid with the network of main roads as of 1913, which we digitized from historical maps published by the Touring Club Italiano (1913). For all cells intersected by a main road segment, we assign the minimum cost value (that is, 1). Hence, the simulated paths between two stage locations should, *ceteris paribus*, closely follow the layout of existing roads. Finally, we generate the route between each pair of consecutive stage locations using the QGIS Least-Cost Path tool, which identifies the optimal path between two locations by minimizing the accumulated cost of travel across the grid based on TRI values.

[42] The data are available on the Copernicus website.