Volume 11, Number 1, 2024

REGION

The Journal of ERSA Powered by WU

Table of Contents

Articles

Funded by

ersa WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS FWF

Articles

REGION
The Journal of ERSA
Powered by WU

Volume 11, Number 1, 2024, 1–26
DOI: 10.18335/region.v11i1.466

journal homepage: region.ersa.org
ISSN: 2409-5370

Labour Market Effects of Trade in a Small Open Economy

Agnes Kügler[1], Klaus Friesenbichler[1], Cornelius Hirsch[1]

[1] Austrian Institute of Economic Research, Vienna, Austria

Received: 6 December 2022/Accepted: 15 December 2023

Abstract. Austria is a small open economy that in the last decades underwent two different waves of increasing trade integration: one with Eastern Europe and one with China. Drawing on trade theory, this paper studies the effects of increases in trade with China and Eastern Europe on labour market dynamics in Austrian NUTS-4 regions for two ten-year periods between 1995 and 2015. Given the limited data available, the current analysis could not identify significant effects on aggregate labour dynamics neither for rising imports from Eastern Europe or China, nor for rising exports to Eastern Europe. However, there is weak evidence that exports to China have facilitated employment growth, especially in high quality segments. Overall, these results add a cautious perspective to the discussion of import competition.

1 Introduction

Over the last two decades, industrialised countries have experienced a strong increase in trade with China. At the same time, trade relations between Western and Eastern European countries have grown substantially. The potential negative impact of the new competitors on economic performance has become a major concern in Western economies. This paper studies Austria, a small open economy, that has naturally served as a docking point for Eastern European countries because of its historical ties and geographical proximity. Its trade relations with Eastern Europe are at least as important as its trade with China. Eastern Europe is economically integrated with Austria in the Central European "manufacturing core" (Friesenbichler et al. 2018, Stehrer, Stöllinger 2015) where firms compete on a regulatory level playing field (Böheim, Friesenbichler 2016, Hölscher, Stephan 2009) and labour-intensive, low-cost segments have, especially since the -1990s, moved to Eastern Europe which changed the competitive positioning.

This paper asks whether Austria's local labour markets have been positively or negatively affected by increased trade with Eastern Europe and China. Its aim is to identify potential employment gains and losses due to increased trade with (i) China and (ii) the group of Eastern European countries. Furthermore, the analysis of trade data specifically focuses on different levels of quality. Vertical differentiation is used as a strategic instrument to alleviate competition (Gabszewicz et al. 1981, Shaked, Sutton 1982, 1987): We argue that specialisation plays a crucial role in the extent of trade competition and its impact on regional local labour markets.

In the remainder, we first provide an overview of the empirical literature and the established theoretical underpinnings for developing the hypotheses that we subsequently test empirically for Austria. Next, we provide descriptive statistics on the developments that set the stage for the regression analysis using a regionalised trade dataset for Austria. Finally, we place the empirical results into a broader context.

2 Previous literature and conjectures

2.1 An overview of the empirical literature

There has been much debate in the empirical literature about the effects of increased imports on employment, especially from low-wage countries. While early studies, such as Grossman (1982), concluded that trade had little impact on US manufacturing employment, later literature showed that international trade flows play an important role for domestic labour markets. Bernard et al. found that rising imports from low-wage countries affected the reallocation of manufacturing within and across industries leading to lower employment growth in the US (Bernard et al. 2006). This negative impact of imports was particularly pronounced for low-skilled workers. Other segments of manufacturing may have even grown in response to strong international demand for US exports (Sachs et al. 1994)[1].

Another set of studies focuses on the impact of trade liberalisation, that is, policy changes. In the US, a strong relationship has been found between the decline in manufacturing employment in the early 2000s and the US granting of permanent normal trade relations with China, which prevented tariff increases (Pierce, Schott 2016). The impact of liberalisation on regional labour market outcomes has been mixed, depending on the magnitude of the tariff cuts (Kovak 2013), with long-term effects on wages more pronounced than short-term effects (Dix-Carneiro, Kovak 2017).

The rise of China in world trade motivated the analysis by Autor et al. (2013) who focused on the impact of trade competition from a low-wage country on regional US labour markets. Since regions differ in terms of specialisation and productivity, the impact of import competition can vary substantially across labour markets. The authors' diagnosis was straightforward: Between 1990 and 2007, rising imports from China caused higher unemployment and lower wages in regional labour markets that host industries exposed to import competing (Autor et al. 2013). The effects were larger in labour markets with more workers with less than a college education (Autor et al. 2015). The trade effects are stronger than the employment shifts caused by technological change. If industries facing high import competition in a local labour market contract, some other industries in the same region might expand and offset the negative employment effects. No significant employment gains in unexposed industries were found for the US (Acemoglu et al. 2016, Autor et al. 2016). Insignificant results for EU-regions have been reported by Hoelzl (2021), but based on very short time series.

In Europe, Chinese import growth is the main driver of international trade. The economic integration of Eastern European economies into the EU was another important component. However, these developments took place in slightly different phases. While China's rapid export growth took place in the 2000s, the Association Agreements (AA) with Eastern European countries were already in place in the mid-1990s. These led to an intensification of trade relations between Eastern Europe and the EU, in particular the industrialised countries of Central Europe. Several analyses of the impact of import competition based on European data support the findings for the US. Import penetration led to a reduction in manufacturing employment in Norway and a reallocation of labour from manufacturing to other sectors in Spain (Balsvik et al. 2015, Donoso et al. 2015). Dauth et al. (2014) examine the impact of the increased trade between Germany and 'the East' (i.e., China and Eastern Europe) on German local labour markets from the late 1980s onwards. There have been substantial job losses in German regions specialised in import-exposed industries. In contrast to the US, these losses have been more than offset by employment gains in regions specialising in export-oriented industries, driven mostly by the rise of Eastern Europe. These results contrast with the findings for the

[1] How import competition affects domestic firms also depends on the specific firm characteristics, such as firm size. Large European firms are found to be more sensitive to trade shocks from low-cost countries while small firms are more susceptible to increasing trade with other high-income countries (Colantone et al. 2014).

US, where offsetting employment gains in other industries have not materialised (Autor et al. 2016).

2.2 Conceptual underpinnings

International trade theory has undergone numerous developments, from classical to the neoclassical, new trade theories, and new classical trade theory. The classical theories are represented as country-based theories. Neoclassical International Trade Theory, often known as factor endowment theory, is represented by the model of Heckscher and Ohlin (HO), two Swedish economists, who developed the Factor Endowment Theory in the 1930s to replace Ricardo's theory of comparative advantage (Ricardo 1817) with numerous components. According to the HO Model international trade is mainly caused by the differential factor endowments of countries. It is advantageous to expose a country to international trade and allow its economy to specialise in accordance with its respective endowments. This typically is the core argument in favour for trade liberalisation and based on the HO model.

In the long run, however, the factor prices might not only be relatively, but also absolutely identical in both countries (Samuelson 1948). This 'factor price equalisation theorem' has often been the basis of opponents of free trade agreements who feared that wages in a high-wage country would fall to the level of the low-wage country, leading to a deterioration of real incomes in the high-wage country. However, due to cross-country differences in factor quality, technology and output prices factor price convergence has hardly been observed. In contrast, empirical evidence suggests a positive correlation between labour productivity and wages (Lam 2015).

At the regional level, the 'export base theory' assumes perfect elasticity of input supply and export demand and suggests that regional output and growth is determined by exogenous demand for a region's exports. Regional growth is positively affected by exports directly. Yet, regional income growth driven by rising export also increases demand for local products, which in turn further stimulates regional income growth. Some regions have developed an export base for manufactured goods due to spatial advantages, but this is not sufficient to ensure sustainable growth. Much of the "secondary" and "tertiary" industry will depend on the success of the export base. This resident industry will ensure that the export base continues to expand as a region develops (North 1955). An increase in demand for a region's exports leads to growing regional income and increased investment not only in the export industry, but also in various other types of economic activity, some of which become new export industries. As a result, the export base of the region tends to become more diversified. This also means that growth in the individual regions tends to be uneven. Ultimately, this model is likely to lead to more equal per capita incomes and a greater dispersion of output with long-term factor mobility.

In contrast, endogenous growth theory emphasises the importance of innovation and technological change. Economic growth depends on the rate of innovation, which in turn is affected by market competition, private investment in R&D, the protection of property rights and patents, and investment in human capital (education and training). Free trade policies can affect long-term growth by accelerating technological change (Grossman, Helpman 1990). Due to spillover effects associated with new technologies, endogenous growth theory suggests that different growth patterns can emerge from specialisation in different types of exports. While all regions can benefit from export growth, regions specialising in goods and services with greater spillover potential tend to grow faster than other regions (Leichenko 2000). However, Grossman, Helpman (1990) suggest that trade protection could accelerate growth if it shifts resources to manufacturing rather than R&D in countries that do not have a comparative advantage in R&D. Overall, from the perspective of endogenous growth theory, the effect of trade policy on long-term growth is an empirical question.

International trade and R&D provide opportunities for knowledge transfer through exports, but at the same time increase potential competition from imports. Firms in different industries show considerable heterogeneity in their ability to respond successfully to increasing trade openness (Chung, Alcácer 2002, Feinberg, Gupta 2004). The effect of exports and imports differs depending on how industries are positioned relative to the global leader (Sakakibara, Porter 2001, Winston Smith 2014). According to modern trade theory, domestic firms producing products that can be easily replaced by inexpensive imports exit the market (Melitz, Ottaviano 2008). In contrast, more productive, technologically sophisticated firms avoid import competition through vertical upgrading. At the industry level, performance improves due to this sorting effect, but aggregate employment might decline due to firms' exit or the relocation of less productive firms to other, less competitive industries (Bernard et al. 2006).

Emerging economies, such as China, have benefited from the outsourcing of production by competitive, multinational firms to low-cost countries. Those multinational firms moved their know-how along with the production. This has reduced frictions in otherwise lengthy industrialisation processes. World market leaders are not necessarily distinguished by price-competitiveness, but above all by the technological content and quality of their goods and services (Baldwin 2016). As technological competencies in China have expanded quickly Chinese firms are increasingly competing on a low-wage, high-tech basis. Following the trade literature, the regional effects of increasing trade with China on domestic industrial performance and employment are contingent on the nature of trade. If a Chinese industry consists of leading firms that are internationally competitive due to their low price, know-how and the technological complexity of their product portfolio, increased import competition from China is expected to have a negative impact on the employment level of the domestic industry. If, on the other hand, the Chinese industry can produce at lower costs, but does not provide the same technological content as domestic firms, competitors can pursue a differentiation strategy by offering different quality levels and therefore escape direct competition (Shaked, Sutton 1982). In this case, increased imports of Chinese products are unlikely to have a negative impact on employment. Yet, quality differentiation can be beneficial for both domestic and foreign industries. This positive effect is more likely when domestic and foreign industries are located at different points in the value chain.

2.3 Conjectures

These considerations suggest differences in the nature of trade that induce different effects on regional labour markets. This leads us to the following conjectures:

Conjecture I: The aggregate effects of exports to Eastern Europe and China on regional labour markets are positive.

Conjecture II: The overall impact of imports from Eastern Europe and China on regional labour markets depends on their quality. The effect is negative for high-quality imports, while the effect of low-quality imports is negligible.

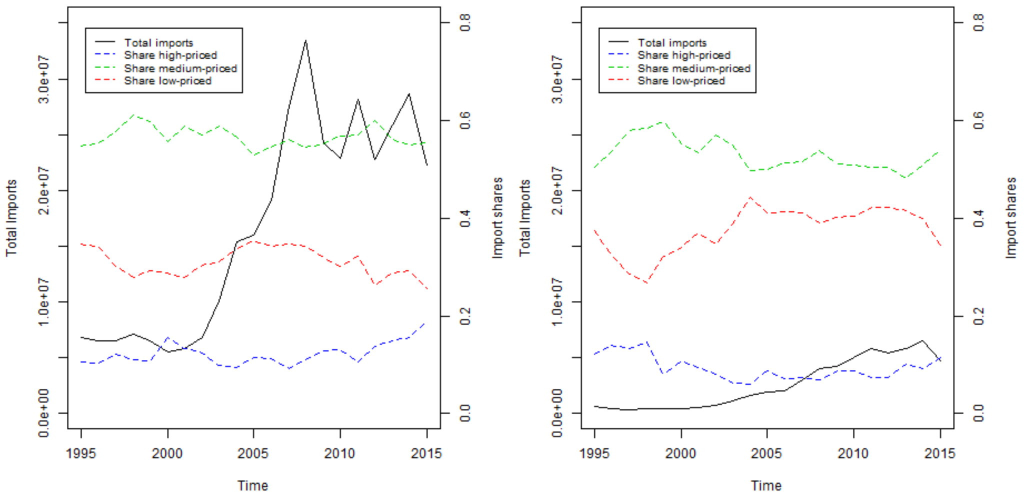
3 Rising trade with Central and Eastern Europe and China

More than the US economy, European countries have been affected by the increase in trade with Central and Eastern European economies, especially in run-up to and after the enlargement of the European Union in 2004, when eight countries (Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia, and Slovenia) became EU members[2]. For Austria, a small open economy bordering Eastern Europe, trade with Eastern Europe has played a particularly important role in recent decades. Imports to and exports from Eastern Europe have grown exponentially since the late 1990s. The increase in trade value between Austria and Eastern Europe over time has been far greater than that of China (see Figure 1 and Figure 2).

The import values from Eastern Europe to Austria show a differentiated structure in terms of quality measured by prices at the product level (see Figure 1)[3]. By far the highest share (2010-2015: 58%) of imports from the Eastern European countries can be classified as medium quality, every fourth imported product is low-quality (2010-2015: 26%) and high-quality imports have the smallest share (2010-2015: 15%). However,

[2] In addition, two Mediterranean countries, Malta, and Cyprus, joined the EU in 2004.

[3] See section Data for the construction of price segments.



(a) Imports from Eastern Europe

(b) Imports from China

Source: BACI, Eurostat price index at NACE-2-digit (2010=100), own calculations.
Notes: In thousand EUR, trade in services not included.

Figure 1: Imports to Austria from Eastern Europe and China between 1995 and 2015

the share of high-quality imports from the Eastern European countries has increased considerably since 2005. In contrast, Austria's share of middle-quality imports (2010-2015: 51%) from China and its share of low-quality imports (2010-2015: 40%) are much more comparable in size. However, the share of high-quality products from China in total imports is still very low (2010-2015: 9%) and has hardly increased over time.

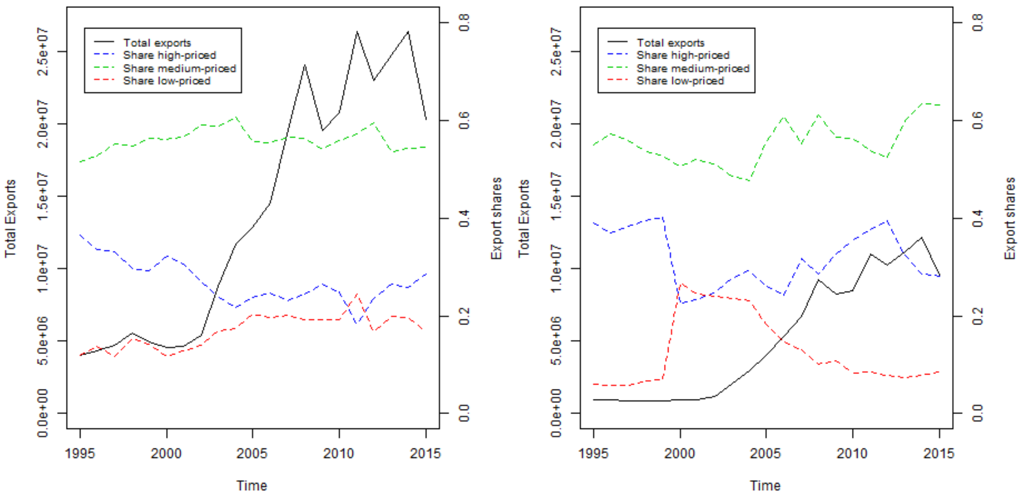
Eastern European economies, but also China, have been proven to be important target markets for Austrian firms. Export dynamics to both countries have gathered pace since the early 2000s (see Figure 2). However, there are differences in the quality of the goods exported to China versus Eastern Europe. Nearly two out of three products exported from Austria to Eastern Europe or China are in the medium price range[4]. However, there are differences with respect to the rest, (i.e., either low- or high-quality goods). For China, more than half of the remainder is comprised of high-quality goods (2010 – 2015: 33%), while the average share of high-quality exports to Eastern Europe is much lower (2010-2015: 24%).

Eastern Europe is a more important trading partner for Austria than China, although trade dynamics with both partners have increased noticeably since the mid-2000s. Most products imported from and exported to Eastern Europe and China are of medium quality (indicated by medium prices). Austria imports a larger share of high-price products from Eastern Europe than from China. Conversely, a higher share of exports to China consists of high-quality products compared to Eastern Europe. Cognisant of this evidence, we ask whether growing trade with these countries has affected Austrian local labour markets. We also study if and to what extent these effects differ according to the quality composition of the traded goods.

4 Empirical Approach and Identification Strategy

To estimate the effects of trade on local labour markets, we first define measures of import and export competition across labour market districts. Following previous literature (Autor et al. 2013, Dauth et al. 2014), our regionalised measures of the change of import and export competition ΔIC_{it}^{C} and ΔEC_{it}^{C} are calculated as:

[4] On average, between 2010 and 2015 about 60% of exports to Eastern Europe and 58% of exports to China consist of medium-priced products.



(a) Exports to Eastern Europe (b) Exports to China

Source: BACI, Eurostat price index at NACE-2-digit (2010=100), own calculations.
Notes: In thousand EUR, trade in services not included.

Figure 2: Exports from Austria to Eastern Europe and China between 1995 and 2015

$$\Delta IC_{it}^{C} = \sum_{j} \frac{E_{ijt}}{E_{jt}} \frac{\Delta IMP_{jt}^{AT\leftarrow C}}{E_{it}} \tag{1}$$

$$\Delta EC_{it}^{C} = \sum_{j} \frac{E_{ijt}}{E_{jt}} \frac{\Delta EXP_{jt}^{AT\rightarrow C}}{E_{it}} \tag{2}$$

where E_{ijt} is the number of employees in region i, industry j at period t, E_{jt} is the aggregate number of employees in industry j, and analogously E_{it} is the aggregate manufacturing employment in region i at time t. $\Delta IMP_{jt}^{AT\leftarrow C}$ and $\Delta EXP_{jt}^{AT\rightarrow C}$ are the changes in industry-specific Austrian imports from and exports to countries (C) such as China or Eastern Europe in real monetary terms (EUR) between time periods t and $t+1$. Thus, ΔIC_{it}^{C} and ΔEC_{it}^{C} capture the potential increase in import and export competition of an Austrian labour market district given its initial sectoral employment structure, since it distributes the national change in sectoral imports among the individual regions according to their shares in national sectoral employment[5].

Our basic regression specification to estimate employment effects at the regional level can be written as:

$$\Delta Emp_{it} = \alpha_0 + \beta_1 \Delta IC_{it}^{C} + \beta_2 \Delta EC_{it}^{C} + X_{it}^{'}\beta_3 + \epsilon_{it}, \tag{3}$$

where $t \in \{1995, 2005\}$, Δ indicates the change between the two 10-year sub-periods 1995-2005 and 2005-2015, and $i \in \{1, \ldots, 85\}$ represents the labour market districts in Austria. ΔEmp_{it} is the 10-year change in the share of manufacturing employment in a region's population in percentage points. $X_{it}^{'}$ is a set of start-of-period control variables varying over regions, such as the share of female workers and the share of ICT specialists employed in a given labour market district. ϵ_{it} depicts the error term, that is clustered at the level of labour market districts to account for spatial or serial correlation.

In a next step, we identify possible differences in the effects of import and export competition in terms of the quality levels of the traded products, focusing on high and low quality for the sake of brevity. Instead of solely looking at the total imports from

[5] To demonstrate the regional heterogeneity over time, Figure A.3 to Figure A.6 in the Appendix show the changes of manufacturing employment shares in Austrian labour markets between 1995 and 2005 and between 2005 and 2015.

and exports to Eastern Europe and China, we use the high-quality imports and exports to calculate high-quality trade competition measures:

$$\Delta IC_{it}^{C,high} = \sum_j \frac{E_{ijt}}{E_{jt}} \frac{\Delta IMP_{high,jt}^{AT\leftarrow C}}{E_{it}} \tag{4}$$

$$\Delta EC_{it}^{C,high} = \sum_j \frac{E_{ijt}}{E_{jt}} \frac{\Delta EXP_{high,jt}^{AT\rightarrow C}}{E_{it}}. \tag{5}$$

Our specification changes to

$$\Delta Emp_{it} = \alpha + \beta_1 \Delta IC_{it}^{C,high} + \beta_2 \Delta EC_{it}^{C,high} + X_{it}^{'}\beta_3 + \epsilon_{it}, \tag{6}$$

The same approach is used to estimate the effects of low-quality imports and export from and to Eastern Europe and China on regional employment in Austria. The effects for trade with China are estimated separately from the effects with Eastern Europe. Trade with Eastern Europe is defined as imports and exports for the country group consisting of Bulgaria, Czech Republic, Hungary, Poland, Romania, Slovakia, Slovenia, Estonia, Latvia, and Lithuania.

Changes in trade volumes could be the result of country- or region-specific demand shocks. Regional employment as well as imports might be positively correlated with unobserved shocks in Austrian product demand. In other words, changes in local labour markets may be the result of other developments than increasing trade with China and Eastern Europe. To identify the causal effect of increasing trade with China and Eastern Europe and account for potential endogeneity of Austrian trade exposure an instrumental variable approach (2SLS) is employed.

Following the approaches used by other papers on import competition (Autor et al. 2013, Bloom et al. 2019, Dauth et al. 2014), we use the composition and growth of Chinese and Eastern European imports from and exports to eight other high-income countries where no significant correlation between demand and supply shocks with Austria is expected. The eight developed non-Euro countries used for the instrumental variable approach are Australia, Canada, Japan, Norway, New Zealand, Sweden, Singapore, and the United Kingdom.

Moreover, to avoid issues in terms of measurement errors and reversed causality because of anticipated future trade competition we use 5-year lagged sectoral employment shares to calculate the instruments (see equations (7) and (8)).

$$\Delta \text{Inst.}IC_{it}^{C} = \sum_j \frac{E_{ijt-5}}{E_{jt-5}} \frac{\Delta IMP_{jt}^{\text{IV - Ctry}\leftarrow C}}{E_{it-5}} \tag{7}$$

$$\Delta \text{Inst.}EC_{it}^{C} = \sum_j \frac{E_{ijt-5}}{E_{jt-5}} \frac{\Delta EXP_{jt}^{\text{IV - Ctry}\rightarrow C}}{E_{it-5}}, \tag{8}$$

5 Data

We confine our analysis to the manufacturing sector. In the period analysed, Austria's imports and exports are dominated by goods rather than services (Reinstaller, Friesenbichler 2020). This may explain why manufacturing still is a very important sector in Austria compared to other high-income countries in Europe (Eurostat 2020). The trade data are obtained from BACI, which is a harmonised trade data set containing information on imports and exports (Gaulier, Zignago 2010). BACI provides information on the quantity of each traded product line at the HS92 6-digit level. However, BACI does not contain industry information. To match the trade data with the industry classification (NACE Rev. 2., 4-digit), we recode HS92 6-digit data to HS02, for which a NACE Rev. 1 correspondence table is available, which again can be transformed into NACE Rev. 2 data at the four-digit level. Unit values are obtained by dividing the export values by the corresponding quantities[6]. For each year and each target market, (i.e., NACE 4-digit industries in different countries), these unit values are aggregated.

We study the impact of trade on regional labour markets. Unfortunately, trade information at a more disaggregated, regional level is unavailable. Hence, we regionalise trade flows according to equations (1) and (2). Against the backdrop of the trade-induced structural change discussion, we further use unit values as a proxy for product quality (Peneder 1999). Especially for trade relations with catching-up economies, product quality is a distinguishing characteristic and a vertical differentiation is a common reaction to competitive dynamics (Hombert, Matray 2018). We use unit values to divide trade flows into high-, and low-quality segments. The trade flows that belong to the upper 25% of the unit values are classified as high-quality exports. Analogously, the lowest 25% of the unit values are classified as low-quality exports. Since sectoral trade flows might be affected by outliers that increase volatility, we use a three-year average (1995-1997, 2004-2006 and 2013-2015) of imports and exports to determine structural differences between the two 10-year periods of interest (1995-2005 and 2005-2015)[7]. The analysis is restricted to trade in manufacturing goods.

The regional employment data are provided by the Federation of Austrian Social Insurance Institutions (‘Hauptverband der österreichischen Sozialversicherungsträger’) and are available for all industries at NACE Rev.2 4-digit level covering 85 different labour market areas (‘Arbeitsmarktbezirk’) in Austria[8]. Depending on the labour market and time, the changes in manufacturing shares vary significantly[9]. The changes in manufacturing shares do not follow a general trend across regions. The data rather show an increase in the manufacturing share in some of Austria’s labour market districts, while others experience a decrease.

Over and above trade exposure, technological change may equally affect labour market dynamics. The literature has discussed two interrelated phenomena for which we control: automation of tasks and the rise in ICT. The effects of increasing automation on the labour force are controversially discussed (Arntz et al. 2016, Bowles 2014, Frey, Osborne 2017). According to Frey, Osborne (2017), 47% of jobs in the US are potentially at high risk of automation. Bowles (2014) transferred this approach to EU countries and calculated that in Austria more than half of all jobs could be affected by automation. However, rather than entire occupations, specific job tasks might be replaced, supported, or created. Using the task-based approach, an OECD Working Paper (Arntz et al. 2016) found that in Austria 12% of employees work in jobs with a high risk of automation. Manual routine jobs are decreasing, and abstract non-routine jobs are increasing in importance (Hölzl et al. 2019). We therefore control for the share of people employed in jobs mainly characterised by routine tasks at the labour-market district level (Peneder et al. 2016)[10].

The changing task structure has been linked to digitalisation (Hölzl et al. 2019). Recent results for Austria show that more ICT intensive economic structures positively affect firm-growth dynamics, which again have been linked to higher employment growth (Friesenbichler, Hölzl 2020). To capture the role of ICT; we apply an taxonomy (Peneder 2020) and compute the number of ICT professionals in different NACE 4-digit industries. We also use the share of regional employment in ICT-intensive industries as another control variable.

In addition, we control for the share of female employees in manufacturing. Since women working in manufacturing are more likely to have low-wage jobs, ceterus paribus,

[6] To exclude measurement errors the unit values are filtered using the filtering method proposed by Gaulier et al. (2008) for the price index calculation.

[7] We use a price index (2010=100) provided by Eurostat at NACE-2-digit level to gain real trade volumes.

[8] Some NACE 4-digit industries have been excluded from the analysis due to their lack of competition, such as mining support service activities (0900) or the postal activities under universal service obligation (5310).

[9] To demonstrate the regional heterogeneity over time, Figure A.1 and Figure A.2 in the Appendix show the change in import and export competition with Eastern Europe and China according to equations (1) to (5) across Austrian regions.

[10] The shares of manual routine and cognitive routine activities are combined into routine tasks per industry.

female workers are expected be more affected by trade shocks than their male colleagues (Autor et al. 2016). Table A.1 in the Appendix presents the summary statistics of all variables used in our regressions[11].

5.1 Data limitations

Employment data are provided by the Federation of Austrian Social Insurance Institutions. The dataset contains information on employers and the number of persons employed in the private sector in Austria (i.e., NACE Rev. 2 sectors A to N.), but the analysis is limited to the manufacturing sector. Self-employed and public-sector employees are not considered. At the firm level, the data are based on social security numbers, and contain the number of employees and the industry affiliation. The use of administrative data is preferable to surveys, because in a highly developed country like Austria the data quality can be assumed to be higher in official records. The number of employees is a figure reported to the social security authority instead of relying on recall information. The employment dataset used starts in 1974, but we restrict the sample to the period from 1995 through 2015 to maintain data comparability. Changes in the sector classifications and data coverage make it difficult to use data from before the mid-1990s. We construct a data set that reports annual employment stocks for all private sector firms with at least one employee at a given reference date[12]. 85 different labour market areas ('Arbeitsmarktbezirk') are covered. The analysis is based on the labour market definition used by the Federation of Austrian Social Insurance Institutions and the Public Employment Service Austria (AMS).

This approach has several disadvantages. First, if the effects of increasing trade with Eastern Europe on regional labour markets occurred before 1995 these are not covered in the analysis. Second, looking at the period between 1995 and 2015 implies that due to the introduction of the new industry classification (NACE Rev.2) in 2008/11 the transition from NACE Rev.1 to NACE Rev. 2 took place in the middle of the observation period. We use correspondence tables between NACE Rev.2 and Rev.1 to smooth the data, but for some industries differences in trade between 2005 and 2015 might be due to the NACE transition and thus distort the results. Third, a disadvantage of these administrative data is that they do not provide information on whether entities are enterprises or establishments. The anonymous firm identifiers in the social security files are administrative accounts. It is left at the discretion of the firm whether it chooses to report at the enterprise or the establishment level. A series of plausibility checks have been carried out to ensure that business units are properly defined. Most of the observations are small firms, which are likely to be at the enterprise level, because having one account reduces administrative burdens when reporting social security contributions (Stiglbauer et al. 2003). Moreover, the analysis is limited to manufacturing sector. While we know that the inconsistent use of enterprises and establishments in this dataset is a huge problem in industries such as retail or financial services, we are confident that it is a minor problem in the manufacturing sector. Despite the issues associated with the use of these employment data, there is no alternative for Austria that would provide both the length of time and the possibility of regionalisation.

6 Results

First, Table 1 shows the estimation results of equation (3). In columns (1) to (4) focus on the effects of import and export competition with Eastern Europe on Austrian labour markets, column (5) to (8) focus on import and export competition with China. OLS estimates are presented in columns (1), (3), (5), and (7). The other columns show

[11] The import and export competition measures are calculated according to equations (1) to (5). The maps in Figure A.3 to Figure A.6 depict the geographical distribution of the change in imports and exports from and to Eastern Europe and China in relation to employees according to equations (1) to (5) across Austrian regions. The brighter the region the higher the increase in import exposure between 1995 and 2005 or between 2005 and 2015 (Figure A.3 and Figure A.4). Similarly, the brighter the region the higher the increase in export possibilities over the same periods of time (Figure A.5 and Figure A.6).

[12] The reporting date is 31 December each year.

the 2-stage-least squares (2SLS) estimates using the instruments discussed above. The corresponding first-stage statistics for the quality of the instrumental variable approach is presented at the bottom of the table[13]. Further, we include the Durbin-Wu-Hausman test statistics. Despite previous literature and theory suggesting that import and export competition from Eastern Europe and China are endogenous in our model, the test statistics imply that our measures of import and export competition are exogenous in both models for Eastern Europe and China at a 10% significance level indicating that we can rely on the OLS estimates. Nonetheless, for completeness and comparability with previous literature, we still provide the 2SLS estimation results. All regressions are estimated with regional dummies.

Neither the OLS-regression, nor the instrumental variable estimations hint at any significant relation between changing regionalised import measures and 10-year changes in manufacturing employment in Austrian labour markets between 1995 and 2015. Our regressions suggest a weak correlation between export competition with China and the Austrian regional labour market, indicating that the increase in exports to China might have resulted in an increase in the share of manufacturing employment The coefficients are only significant at 10 percent level, though. In contrast, the results show no statistically significant effect of export competition with Eastern Europe on manufacturing employment. The coefficients of the regional manufacturing share at the starting period are significant and negative indicating that higher starting values of manufacturing shares are related to lower increases in the next ten years. The impacts of neither the share of manufacturing jobs dominated by routine tasks in 1995 and 2005 nor the share of female employees or of ICT-intensive industries in a region are significantly different from zero.

In a second step we test whether these results depend on the quality of imports from and exports to Eastern Europe and China measured by different price segments (high, and low prices). The estimation results of equation (6) are shown in Table 2. Again, columns (1), (3), (5), and (7) show OLS estimation results, while columns (3), (4), (6) and (8) show the results based on 2SLS regressions. In the high-quality segment, no significant effects of changes in import or export competition on regional employment can be observed, neither for Eastern Europe (columns (1) to (4)) nor for China (columns (5) to (8)). Similarly, for the low-quality segment, the results in Table 3 suggest no statistically significant correlation between changing import and export competition with Eastern Europe or China and the 10-year changes in manufacturing employment in Austrian labour markets between 1995 and 2015.

Overall, these regression results suggest that import competition has played only a minor role in labour market changes in Austria between 1995 and 2015. Neither the results for Chinese nor for Eastern European imports allow one to deduce serious effects on regional employment in Austria, regardless of the quality segment. Thus, the results do not support Conjecture II. The effects of export competition seem to be somewhat larger. Conjecture I is partly supported: While no effects of increasing export competition with Eastern Europe can be observed, our results indicate a positive correlation between increasing export competition with China and regional manufacturing employment in Austria.

7 Discussion

Previous results are largely available for countries such as the US or Germany. However, Austrian regions are more homogeneous in terms of their economic development and industry structure than German regions or regions in the United States. The differences in the regional industrial specialisation (e.g., the performance differences between German states in the "East" and "West" which continue to exist after the German reunification, or the "rust belt" as opposed to the ICT-intensive coastal areas in the US) are different from those in a small and rather homogeneous country such as Austria. Thus, the

[13] The Cragg-Donald test statistic suggests that the maximum bias of the 2SLS estimator will be no more than 10% of the bias of OLS at a 5% significance level. The Kleibergen-Paap Wald test statistic suggests that the maximum bias of the 2SLS estimator will be no more than 15% of the bias of OLS.

Table 1: Effects of Import and Export Competition with Eastern Europe and China on Manufacturing Employment in Austrian Labour Markets

	Dependent variable: *10-year change manufacturing employment/working age pop. in %-points*							
	Eastern Europe				China			
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)	OLS (7)	2SLS (8)
Δ Import competition with ... (3-year average)	0.01	-0.04	-0.02	0.02	-0.03	-0.01	-0.06	-0.11
	(0.03)	(0.09)	(0.03)	(0.11)	(0.07)	(0.18)	(0.07)	(0.14)
Δ Export competition with ... (3-year average)	-0.01	0.09	-0.01	0.02	0.05+	0.07+	0.01	0.03
	(0.01)	(0.11)	(0.01)	(0.06)	(0.03)	(0.04)	(0.02)	(0.03)
Share of manufacturing jobs in total employment			-0.33***	-0.31***			-0.33***	-0.33***
			(0.09)	(0.07)			(0.08)	(0.07)
Share of employment in ICT-intensive industries			-0.02	0.00			0.01	0.01
			(0.05)	(0.08)			(0.05)	(0.06)
Share of routine jobs in manufacturing employment			0.09	0.03			0.08	0.07
			(0.19)	(0.15)			(0.17)	(0.12)
Share of female employees in manufacturing			-0.07	-0.08			-0.08	-0.09
			(0.07)	(0.09)			(0.07)	(0.07)
Constant	0.44	-4.06	2.08	2.14	-1.73	-2.64+	1.55	1.09
	(0.92)	(5.34)	(8.28)	(8.25)	(1.14)	(1.56)	(8.04)	(5.78)
Observations	170	170	170	170	170	170	170	170
Cragg-Donald Wald F statistic		11.63		10.13		12.86		9.46
Kleibergen-Paap Wald F statistic		5.47		4.29		5.81		5.29
Wu-Hausman p-value		0.06		0.12		0.29		0.15

Notes: Regional fixed effects included in all regressions. Instruments are based on the composition and growth of Chinese and Eastern Europe imports from and exports to eight other high-income non-Euro countries: Australia, Canada, Japan, Norway, New Zealand, Sweden, Singapore, and the United Kingdom. Clustered standard errors in parenthesis; + p<0.1, * p<0.05, ** p<0.01, *** p<0.001. Stock-Yogo weak ID test critical values: Acceptable level of bias= 10%: 7.03, Acceptable level of bias= 15%: 4.58, Acceptable level of bias= 20%: 3.95 (Stock, Yogo 2005)

Table 2: Effects of Competition in High–quality Imports and Exports from and to Eastern Europe and China on Manufacturing Employment in Austrian Labour Markets

	Dependent variable: *10-year change manufacturing employment/working age pop. in %-points*							
	Eastern Europe				China			
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ Import competition with ... (3-year average)	-0.02	0.28	-0.08	0.09	-0.24	-1.63	-0.40	-4.33
	(0.12)	(0.60)	(0.13)	(0.41)	(0.34)	(2.20)	(0.32)	(5.22)
Δ Export competition with ... (3-year average)	0.04	-0.18	0.01	-0.13	0.16	0.37	0.04	0.58
	(0.07)	(0.46)	(0.06)	(0.32)	(0.11)	(0.34)	(0.09)	(0.66)
Share of manufacturing jobs in total employment			-0.34***	-0.36***			-0.34***	-0.48*
			(0.09)	(0.09)			(0.08)	(0.19)
Share of employment in ICT-intensive industries			-0.01	0.00			-0.01	-0.05
			(0.04)	(0.03)			(0.04)	(0.09)
Share of routine jobs in manufacturing employment			0.09	0.13			0.10	0.29
			(0.18)	(0.16)			(0.16)	(0.30)
Share of female employees in manufacturing			-0.08	-0.09			-0.08	-0.13
			(0.07)	(0.06)			(0.07)	(0.13)
Constant	-0.60	2.56	1.61	2.18	-0.98	-2.07	1.01	-7.24
	(1.08)	(6.66)	(8.07)	(6.43)	(0.82)	(1.76)	(7.87)	(12.90)
Observations	170	170	170	170	170	170	170	170
Cragg-Donald Wald F statistic		3.79		2.50		3.79		3.75
Kleibergen-Paap Wald F statistic		1.27		1.57		0.63		0.72
Wu-Hausman p-value		0.61		0.74		0.37		0.00

Notes: Regional fixed effects included in all regressions. Instruments are based on the composition and growth of Chinese and Eastern Europe imports from and exports to eight other high-income non-Euro countries: Australia, Canada, Japan, Norway, New Zealand, Sweden, Singapore, and the United Kingdom. Clustered standard errors in parenthesis; + p<0.1, * p<0.05, ** p<0.01, *** p<0.001. Stock-Yogo weak ID test critical values: Acceptable level of bias= 10%: 7.03, Acceptable level of bias= 15%: 4.58, Acceptable level of bias= 20%: 3.95 (Stock, Yogo 2005)

Table 3: Effects of Competition in Low–quality Imports and Exports from and to Eastern Europe and China on Manufacturing Employment in Austrian Labour Markets

	Dependent variable: *10-year change manufacturing employment/working age pop. in %-points*							
	Eastern Europe				China			
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ Import competition with ... (3-year average)	-0.02	-0.01	0.03	0.12	-0.09	0.44	-0.16	0.16
	(0.05)	(0.16)	(0.05)	(0.21)	(0.15)	(0.35)	(0.14)	(0.34)
Δ Export competition with ... (3-year average)	0.06	0.38	0.06	0.25	-0.05	0.04	-0.01	0.03
	(0.08)	(0.32)	(0.06)	(0.19)	(0.13)	(0.05)	(0.06)	(0.04)
Share of manufacturing jobs in total employment			-0.33***	-0.34***			-0.33***	-0.32***
			(0.09)	(0.08)			(0.08)	(0.07)
Share of employment in ICT-intensive industries			0.01	0.05			0.01	-0.02
			(0.05)	(0.06)			(0.05)	(0.05)
Share of routine jobs in manufacturing employment			0.06	0.02			0.07	0.08
			(0.18)	(0.13)			(0.18)	(0.13)
Share of female employees in manufacturing			-0.08	-0.11			-0.08	-0.05
			(0.07)	(0.07)			(0.06)	(0.07)
Constant	-0.41	-3.11	1.79	1.47	0.88	-1.15	2.37	0.99
	(0.72)	(2.64)	(8.30)	(6.34)	(1.64)	(1.04)	(8.09)	(6.40)
Observations	170	170	170	170	170	170	170	170
Cragg-Donald Wald F statistic		17.34		16.36		12.46		7.36
Kleibergen-Paap Wald F statistic		8.30		6.07		3.29		2.13
Wu-Hausman p-value		0.05		0.07		0.14		0.33

Notes: Regional fixed effects included in all regressions. Instruments are based on the composition and growth of Chinese and Eastern Europe imports from and exports to eight other high-income non-Euro countries: Australia, Canada, Japan, Norway, New Zealand, Sweden, Singapore, and the United Kingdom. Clustered standard errors in parenthesis; + p<0.1, * p<0.05, ** p<0.01, *** p<0.001. Stock-Yogo weak ID test critical values: Acceptable level of bias= 10%: 7.03, Acceptable level of bias= 15%: 4.58, Acceptable level of bias= 20%: 3.95 (Stock, Yogo 2005)

regional analysis of Austria is based on a smaller number of less heterogeneous observations than in the case of the US or Germany. Moreover, Austria's trade integration is deep and some of its major industries like the manufacturing of machinery and transport equipment are important suppliers for industries in Germany and other countries. However, indirect effects of changes in import competition in these customer countries are not covered by the present analysis.

When interpreting the result for Austria, some notable institutional differences should be considered. The Austrian labour market is more rigid than the US labour market (Lithuania Free Market Institute 2019). The Austrian system of 'economic and social partnership' is characterised by a high degree of corporatism, involving economic chambers and trade unions in collective wage-bargaining and the parts of the content of labour market policies. Taken together, these factors may explain why, compared to Germany or the US, there are no large effects of trade competition at the regional level.

The Austrian industry is characterised by a high share of small firms. According to the Structural Business Statistics provided by Statistics Austria, 69.9% of the manufacturing firms employed fewer than ten people in 2015. Only 1.8% of the firms reported more than 250 employees. According to literature, small and medium-sized firms are less sensitive to trade shocks from low-cost countries than large firms (Colantone et al. 2014). Furthermore, small firms tend to compete in more protected niche markets (Porter 1980, Spanos et al. 2004). In other words, replacement effects that underlie the argument may not be present. Imports from Eastern Europe are often complementary products in regional value chains (Friesenbichler et al. 2018, Stehrer, Stöllinger 2015). It is possible that Chinese products are targeting global markets rather than small niche markets. Thus, even though the products are in the same industry class, they could be targeting different demand groups.

Peneder (1999) demonstrates that Austria's industrial pattern is most unusual compared to other high-performing countries. While the US and Germany have been characterised by high shares of technology-driven industries, Austria has a high share of mainstream manufacturing combined with a low share of technology-driven industries. However, compared to other countries these 'traditional' industries in Austria are rather innovative in terms of patents and R&D. Within mainstream manufacturing Austria clearly outperformed countries with comparably high manufacturing shares: 'The fact that in 1997, the labour productivity of total manufacturing in Austria was 46 percent above that of Spain and 69 percent above that of Portugal illustrates that similar patterns of specialisation can still comprise very different kinds of activities' (Peneder 1999, p. 244). This might explain the differences of the impact of trade competition with China on local labour markets between countries with comparable industry structure, such as the labour intensive economy of Spain (Donoso et al. 2015). Also, the insignificant effects of trade competition from Eastern Europe and China on regional Austrian labour markets could be due to specialisation in different market segments within NACE 4-digit industry classes which is not limited to vertical differentiation and therefore not captured by using unit values to identify different quality segments. In other words, the level of industry aggregation (NACE 4-digit) might be still too high to identify any significant effects of import competition on local labour markets.

In addition, the exposure of employment-intensive industries to trade has been rather small. The descriptive statistics show that imports from Eastern Europe and China tend to be in NACE 4-digit industries that are rather small in terms of employment shares[14]. While the growth of imports in these industries is noteworthy, the respective Austrian employment shares are small in 1995 and 2005. The same holds for the industries with the highest growth rates of imports from China between 1995 and 2015[15]. Except for sawing

[14] Figure A.7 and Figure A.8 show the NACE 4-digit sectors with the highest increases in imports from Eastern Europe between 1995, 2005 and 2015. The growth of imports from Eastern Europe was the largest in 'Manufacture of watches and clocks' (2652), 'Manufacture of imitation jewelry and related article' (3213) and 'Precious metals production' (2441) between 1995 and 2005, and in 'Manufacture of cider and other fruit wine' (1103), 'Precious metals production' (2441) and 'Manufacture of ice cream' (1052) between 2005 and 2015.

[15] Figure A.9 and Figure A.10 illustrate that between 1995 and 2005 import growth from China was the highest in 'Sawmilling and planing of wood' (1610), 'Manufacture of musical instrument' (3220)

and planing (NACE 1610: employment share of 1.6% in 1995) and the manufacturing of beer (NACE 1105: employment share of 0.75% in 2005), all industries showed shares in manufacturing employment significantly below 0.5% in 1995 and 2005 (the average employment share of a sector was about 0.5% in both years). Another noteworthy exception is the 'Production of meat and poultry meat products' (1013), which accounted for about 2.2% of all employees in manufacturing in 2005. At the same time, between 2005 and 2015 this industry had the fourth highest change in imports from Eastern Europe and its imports from China have grown drastically. However, in 2015 still 2.14% of manufacturing employees worked in the production of meat and poultry meat product. Similarly, the employment share of sawing and planing was in 2015 at the same level (1.6%) as in 1995. Only employment in the manufacturing of beer show a relative decline from 0.9% in 1995 to 0.4% of total manufacturing employment in 2015.

In contrast, looking at the largest industries in terms of manufacturing employment in Austria shows that in these industries the increase in imports was significantly lower. It ranged from -22% to +483% for imports from Eastern Europe and between -1% and +299% for imports from China[16]. Considering the median change in imports between 1995 and 2005 from Eastern Europe was 125% (2005-2015: 31%) and from China 530% (2005-2015: 217%), the increases in imports in the largest Austrian manufacturing industries, particularly from China, are rather small. Even after removing the most extreme outliers (which are all characterised by extremely high import growth rates in very small industries), the relationship between import growth rates and employment shares is negative, although mostly statistically insignificant. For Chinese imports between 1995 and 2005, there is a significant negative correlation between manufacturing employment shares and import growth. This suggests that the strongest Chinese import growth took place mostly in Austrian industries that were already very small in 1995.

A study for the US finds that internal migration is affected by changes in import competition from China (Greenland et al. 2019). Due to increasing trade with China, population adjustments appear especially dynamic in local labour markets that are most exposed to import competition from China. A decline in regional population growth tends to materialise seven to ten years after the trade-enhancing policy change occurred. It mainly affects young and the less educated. Our dependent variable, ΔEmp_{it}, is calculated as the 10-year change in the share of manufacturing employment in a region's population in percentage points. Thus, our left hand side variable might be affected by a trade-induced change in population growth and causal identification of the effects is confounded. Failure to account for changes in the composition of the labour force may lead to biased estimates of the impact of trade on average outcomes at the level of commuting zones, such as wages or unemployment rates (Greenland et al. 2019, p. 49). This bias is most likely negative in the sense that previous studies might underestimate the effects of import competition from China on US labour markets. Indeed, if regional population growth in Austria was similarly or even more affected by trade competition from Eastern Europe and/or China, this might explain the lack of significant effects of import competition on regional employment to some extent.

8 Conclusions

Austria, a small, industrialised, open economy in Central Europe, has experienced two waves of economic integration in recent decades: one with Eastern Europe and one with China. This paper asked if imports from and exports to both China and Eastern Europe have affected regional labour markets in two ten-year periods: one between 1995 and 2005 and one between 2005 and 2015. Neither increases in imports from nor exports to

and 'Manufacture of jewelry and related article' (3212). From 2005 to 2015 the highest increases were observed in 'Manufacture of homogenised food preparations and dietetic food' (1086), 'Manufacture of beer' (1105) and 'Manufacture of knitted and crocheted hosiery' (1431).

[16] In 1995, the largest Austrian manufacturing sectors having a share in manufacturing employment above 3% were 'Manufacture of other furniture' (31.09), 'Manufacture of bread; manufacture of fresh pastry goods and cake' (1071) and 'Manufacture of basic iron and steel and of ferro-alloy' (2410). In 2005, besides 'Manufacture of other furniture' (31.09), and 'Manufacture of bread; manufacture of fresh pastry goods and cake' (1071), 'Manufacture of motor vehicle' (29.10) had a share of more than 3% of all employees in manufacturing (see Figure A.11 and Figure A.12 in the Appendix).

Eastern Europe have had a significant impact on aggregate labour dynamics. Austria seems to have benefited from the emerging "manufacturing core" in Central and Eastern Europe since 1990 (Stehrer, Stöllinger 2015) and fears of job losses are not visible, at least in aggregate dynamics. The data also show no significant impact from rising imports from China, either. However, there is some evidence that exports to China facilitate employment growth, especially in high-quality segments.

However, competitive pressure, especially from Chinese products, may follow a stage model in which the aggregate effects on the labour market occur last. Before the aggregate effects become visible, there are firm-level effects such as a decline in firms' competitiveness or firm exits in industries competing with China (Branstetter et al. 2019). Prior to firm-level changes in employment stocks, firms react in their strategic behaviour. Austrian firms have been shown to adjust their strategies in response to international competition. Even small changes in exposure to Chinese competition have a significant impact on diversification decisions. Firms exposed to increasing Chinese competition are more likely to diversify their geographical markets, but less likely to diversify their product portfolio or broaden their competence base (Friesenbichler, Reinstaller 2022, 2023). If Chinese growth continues, the pressure of Chinese import competition could eventually be reflected in labour market figures.

Nevertheless, these results put the current discussion about import competition into perspective. Neither increased internationalisation with China nor European economic integration with Eastern Europe has (negatively) affected Austrian labour markets to an extent that would be visible in aggregate terms. This is also true for the control variables. There are labour market dynamics in terms of the demand for tasks, and manual, routine tasks have become rarer. However, this has not affected the aggregate results.

References

Acemoglu D, Autor D, Dorn D, Hanson GH, Price B (2016) Import competition and the great US employment sag of the 2000s. *Journal of Labor Economics* 34. CrossRef

Arntz M, Gregory T, Zierahn U (2016) The risk of automation for jobs in OECD countries: A comparative analysis. OECD working papers, 189. CrossRef

Autor D, Dorn D, Hanson GH (2013) The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review* 103: 2121–2168. CrossRef

Autor DH, Dorn D, Hanson GH (2015) Untangling trade and technology: Evidence from local labour markets. *The Economic Journal* 125: 621–646. CrossRef

Autor DH, Dorn D, Hanson GH (2016) The China shock: Learning from labor market adjustment to large changes in trade. *Annual Review of Economics* 8: 205–240. CrossRef

Baldwin R (2016) *The Great Convergence. Information Technology and the New Globalization.* The Belknap Press of Harvard University Press, Cambridge, MA, London, England. CrossRef

Balsvik R, Jensen S, Salvanes KG (2015) Made in China, sold in Norway: Local labor market effects of an import shock. *Journal of Public Economics* 127: 137–144. CrossRef

Bernard AB, Jensen J, Schott PK (2006) Survival of the best fit: Exposure to low-wage countries and the (uneven) growth of US manufacturing plants. *Journal of International Economics* 68: 219–237. CrossRef

Bloom N, Handley K, Kurmann A, Luck P (2019) The impact of Chinese trade on US employment: The good, the bad, and the apocryphal. 2019 meeting papers 1433, society for economic dynamics

Bowles J (2014) 54% of EU jobs at risk of computerisation. Bruegel blog post, https://www.bruegel.org/blog-post/chart-week-54-eu-jobs-risk-computerisation

Branstetter L, Kovak B, Mauro J, Venancio A (2019) *The China Shock and Employment in Portuguese Firms*, Volume w26252. National Bureau of Economic Research, Cambridge, MA. CrossRef

Böheim MH, Friesenbichler KS (2016) Exporting the competition policy regime of the European Union: Success or failure? Empirical evidence for acceding countries. *Journal of Common Market Studies* 54: 569–582. CrossRef

Chung W, Alcácer J (2002) Knowledge seeking and location choice of foreign direct investment in the United States. *Management Science* 48: 1534–1554. CrossRef

Colantone I, Coucke K, Sleuwaegen L (2014) Low-cost import competition and firm exit: Evidence from the EU. *Industrial and Corporate Change* 24: 131–161. CrossRef

Dauth W, Findeisen S, Suedekum J (2014) The rise of the East and the Far East: German labor markets and trade integration. *Journal of the European Economic Association* 12: 1643–1675. CrossRef

Dix-Carneiro R, Kovak BK (2017) Trade liberalization and regional dynamics. *American Economic Review* 107: 2908–2946. CrossRef

Donoso V, Martín V, Minondo A (2015) Do differences in the exposure to Chinese imports lead to differences in local labour market outcomes? An analysis for Spanish provinces. *Regional Studies* 49: 1746–1764. CrossRef

Eurostat (2020). Manufacturing statistics

Feinberg SE, Gupta AK (2004) Knowledge spillovers and the assignment of R&D responsibilities to foreign subsidiaries. *Strategic Management Journal* 25: 823–845. CrossRef

Frey CB, Osborne MA (2017) The future of employment: How susceptible are jobs to computerisation. *Technological Forecasting and Social Change* 114: 254–280. CrossRef

Friesenbichler K, Hölzl W (2020) High-growth firm shares in Austrian regions: The role of economic structures. *Regional Studies* 54: 1585–1595. CrossRef

Friesenbichler K, Reinstaller A (2022) Do firms facing competitors from emerging markets behave differently? Evidence from Austrian manufacturing firms. *European Business Review* 34: 153–170. CrossRef

Friesenbichler KS, Glocker C, Hölzl W, Kaniovski S, Reinstaller A, Streicher G, Stehrer R, Stöllinger R, Leitner S, Hanzl-Weiss D, Reiter O, Adarov A, Bykova A, Siedschlag I, Ubaldo M, Studnicka Z (2018) Drivers and obstacles to competitiveness in the EU: The role of value chains and the single market. In: *Research carried out for the European Commission, DG GROW within the Framework Service Contract No. ENTR/300/PP/2013/FC-WIFO under the project "Competitiveness drivers and obstacles, intra-EU linkages and European value chains in GVCs"*. European Commission, Brussels. CrossRef

Friesenbichler KS, Reinstaller A (2023) Small and internationalized firms competing with Chinese exporters. *Eurasian Business Review* 13: 167–192. CrossRef

Gabszewicz J, Shaked A, Sutton J, Thisse J (1981) International trade in differentiated products. *International Economic Review*: 527–534. CrossRef

Gaulier G, Martin J, Méjean I, Zignago S (2008) International trade price indices

Gaulier G, Zignago S (2010). Baci: International trade database at the product-level (the 1994-2007 version)

Greenland A, Lopresti J, McHenry P (2019) Import competition and internal migration. *The Review of Economics and Statistics* 101: 44–59. CrossRef

Grossman GM (1982) The employment and wage effects of import competition in the United States. National bureau of economic research. CrossRef

Grossman GM, Helpman E (1990) Trade, innovation, and growth. National bureau of economic research. CrossRef

Hoelzl W (2021) Import competition from China in manufacturing after the financial crisis: Evidence for European regions. Wifo working papers, 622

Hombert J, Matray A (2018) Can innovation help U.S. manufacturing firms escape import competition from China? *The Journal of Finance* 73: 2003–2039. CrossRef

Hölscher J, Stephan J (2009) Competition and antitrust policy in the enlarged European Union: A level playing field? *Journal of Common Market Studies* 47: 863–889. CrossRef

Hölzl W, Bärenthaler-Sieber S, Bock-Schappelwein J, Friesenbichler K, Kügler A, Reinstaller A, Reschenhofer P, Dachs B, Risak M (2019) *Digitalisation in Austria: State of Play and Reform Needs*. European Union, Luxembourg

Kovak BK (2013) Regional effects of trade reform: What is the correct measure of liberalization? *American Economic Review* 103: 1960–1976. CrossRef

Lam TD (2015) A review of modern international trade theories. *American Journal of Economics* 1

Leichenko RM (2000) Exports, employment, and production: A causal assessment of U.S. states and regions. *Economic Geography* 76. CrossRef

Lithuania Free Market Institute (2019) *Employment Flexibility Index 2020*. EU and OECD Countries

Melitz MJ, Ottaviano GI (2008) Market size, trade, and productivity. *The Review of Economic Studies* 75: 295–316. CrossRef

North DC (1955) Location theory and regional economic growth. *Journal of Political Economy* 63: 243–258. CrossRef

Peneder M (1999) The Austrian paradox: ‘Old’ structures but high performance? *Austrian Economic Quarterly* 4: 239–247

Peneder M (2020) Eine neue Taxonomie zur Gliederung von Branchen nach ihrer IKT-Intensität. *WIFO Monatsberichte* 93

Peneder M, Bock-Schappelwein J, Firgo M, Fritz O, Streicher G (2016) *Österreich im Wandel der Digitalisierung*. WIFO, Österreichisches Institut für Wirtschaftsforschung

Pierce JR, Schott PK (2016) The surprisingly swift decline of US manufacturing employment. *American Economic Review* 106: 1632–1662. CrossRef

Porter ME (1980) Competitive strategy: Techniques for analyzing industries and competitors. In: Porter ME (ed), *Competitive Strategy*. Free, New York

Reinstaller A, Friesenbichler KS (2020) *“Better Exports” – Technologie-, Qualitätsaspekte und Innovation des Österreichischen Außenhandels im Kontext der Digitalisierung*. WIFO, Wien

Ricardo D (1817) *The Works and Correspondence of David Ricardo*, Volume 1. Principles of Political Economy and Taxation. Online Library of Liberty

Sachs JD, Shatz HJ, Deardorff A, Hall RE (1994) Trade and jobs in US manufacturing. *Brookings Papers on Economic Activity* 1994: 1–84. CrossRef

Sakakibara M, Porter ME (2001) Competing at home to win abroad: Evidence from Japanese industry. *Review of Economics and Statistics* 83: 310–322. CrossRef

Samuelson PA (1948) International trade and the equalisation of factor prices. *The Economic Journal* 58: 163–184. CrossRef

Shaked A, Sutton J (1982) Relaxing price competition through product differentiation. *The Review of Economic Studies* 3–13. CrossRef

Shaked A, Sutton J (1987) Product differentiation and industrial structure. *The Journal of Industrial Economics*: 131–146. CrossRef

Spanos YE, Zaralis G, Lioukas S (2004) Strategy and industry effects on profitability: Evidence from Greece. *Strategic Management Journal* 25: 139–165. CrossRef

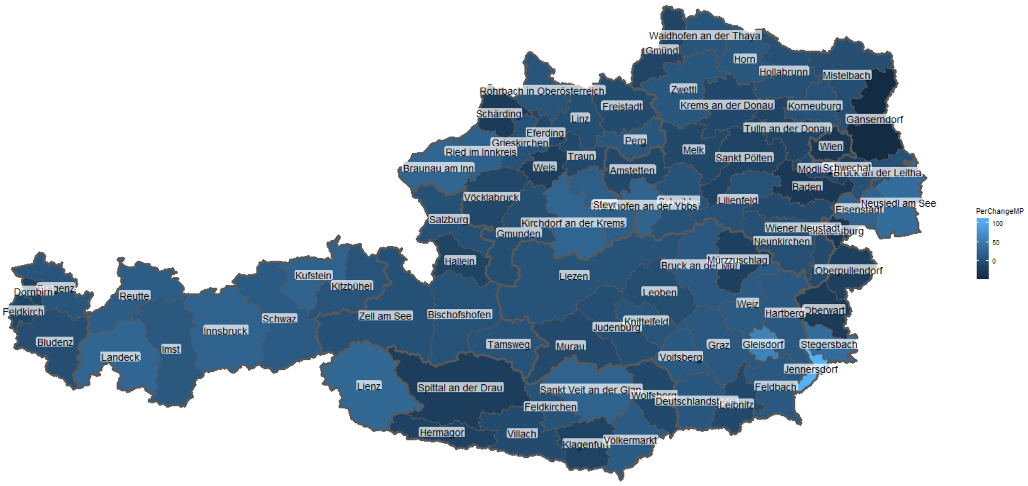
Stehrer R, Stöllinger R (2015) *The Central European Manufacturing Core: What Is Driving Regional Production Sharing?* FIW Research Reports, Vienna. Issue: 2014/15-02

Stiglbauer AM, Stahl F, Winter-Ebmer R, Zweimüller J (2003) Job creation and job destruction in a regulated labor market: The case of Austria. *Empirica* 30: 127–148. CrossRef

Stock JH, Yogo M (2005) Asymptotic distributions of instrumental variables statistics with many instruments. In: Andrews DW, Stock JH (eds), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg.* Cambridge University Press, Cambridge, MA, 109–120. CrossRef

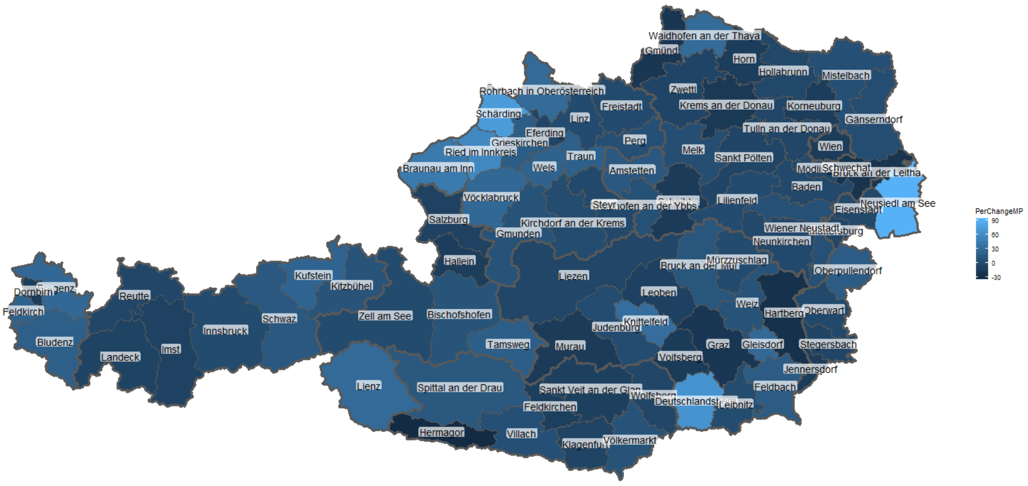
Winston Smith S (2014) Follow me to the innovation frontier? Leaders, laggards, and the differential effects of imports and exports on technological innovation. *Journal of International Business Studies* 45: 248–274. CrossRef

A Appendix:



Notes: The brighter the region the higher the increase in manufacturing employment shares between 1995 and 2005.

Figure A.1: Changes of manufacturing employment shares in Austrian labour markets ('Arbeitsmarktbezirke') between 1995 and 2005 (in %)



Notes: The brighter the region the higher the increase in manufacturing employment shares between 2005 and 2015

Figure A.2: Changes of manufacturing employment shares in Austrian labour markets ('Arbeitsmarktbezirke') between 2005 and 2015 (in %)

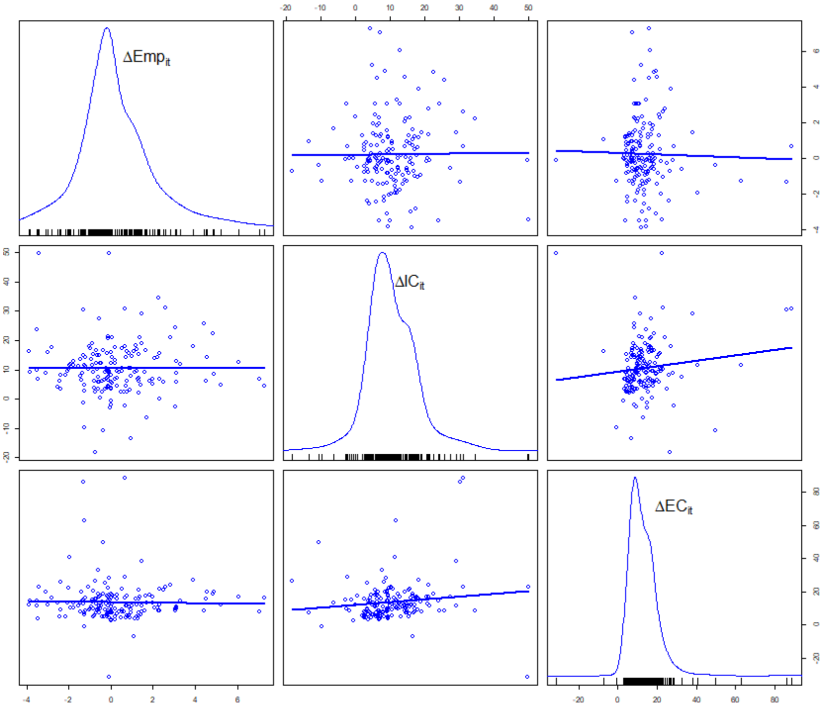


Figure A.3: Scatterplot of the main variables for Eastern Europe

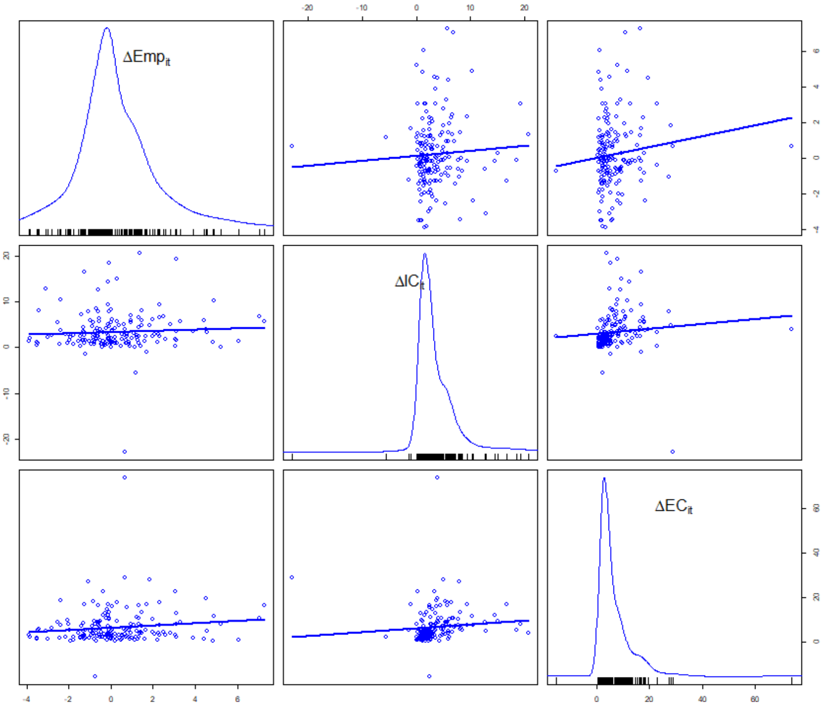
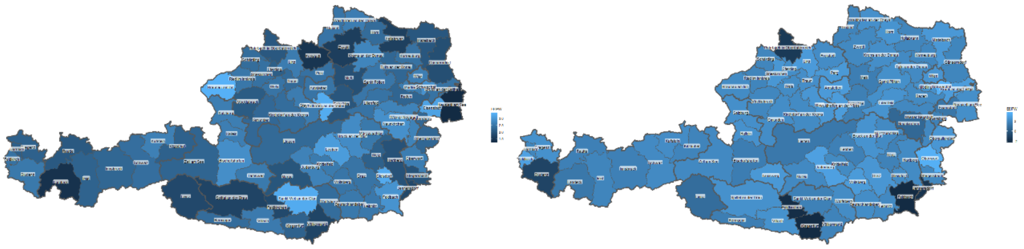


Figure A.4: Scatterplot of the main variables for China

Table A.1: Summary statistics

Variables	N	Mean	Std. Dev.	Min	Max
Change of Manufacturing employment shares in percentage points (population based)	170	0.25	1.95	-3.91	7.29
Change in import competition with Eastern Europe	170	10.69	9.03	-18.18	49.77
Change in export competition with Eastern Europe	170	13.78	12.09	-31.33	88.70
Change in high-quality import competition with Eastern Europe	170	2.44	2.47	-2.37	13.37
Change in high-quality export competition with Eastern Europe	170	3.86	4.12	-0.46	34.98
Change in low-quality import competition with Eastern Europe	170	1.95	3.39	-6.74	15.25
Change in low-quality export competition with Eastern Europe	170	2.26	2.70	-6.39	20.65
Change in import competition with China	170	3.49	4.28	-22.89	20.67
Change in export competition with China	170	6.86	7.85	-15.15	73.91
Change in high-quality import competition with China	170	0.41	0.82	-1.15	9.57
Change in high-quality export competition with China	170	2.28	2.32	-1.60	11.74
Change in low-quality import competition with China	170	1.31	1.62	-8.26	9.85
Change in low-quality export competition with China	170	0.56	2.18	-3.33	26.25
Share of manufacturing jobs in total employment	170	27.18	9.89	9.10	58.15
Share of ICT-intensive industries in manufacturing employment	170	27.90	15.05	1.27	68.20
Share of routine jobs in manufacturing employment	170	47.82	3.22	38.77	56.47
Share of female employees in manufacturing employment	170	27.24	6.48	13.05	59.65
Change in instrument for import competition with Eastern Europe	170	40.49	41.13	2.54	284.32
Change in instrument for export competition with Eastern Europe	170	16.10	16.82	-60.52	98.87
Change in instrument for high-quality import competition with Eastern Europe	170	12.17	17.85	-0.66	192.09
Change in instrument for high-quality export competition with Eastern Europe	170	1.72	2.44	-3.89	22.26
Change in instrument for low-quality import competition with Eastern Europe	170	7.08	8.99	-24.25	70.01
Change in instrument for low-quality export competition with Eastern Europe	170	5.53	7.39	-33.89	38.65
Change in instrument for import competition with China	170	286.28	406.09	-60.10	4422.10
Change in instrument for export competition with China	170	138.47	173.09	-466.25	1019.77
Change in instrument for high-quality import competition with China	170	82.45	86.54	-15.65	867.99
Change in instrument for high-quality export competition with China	170	14.09	21.63	-9.15	122.68
Change in instrument for low-quality import competition with China	170	33.96	134.30	-103.72	1612.11
Change in instrument for low-quality export competition with China	170	50.08	97.90	-277.19	1044.91

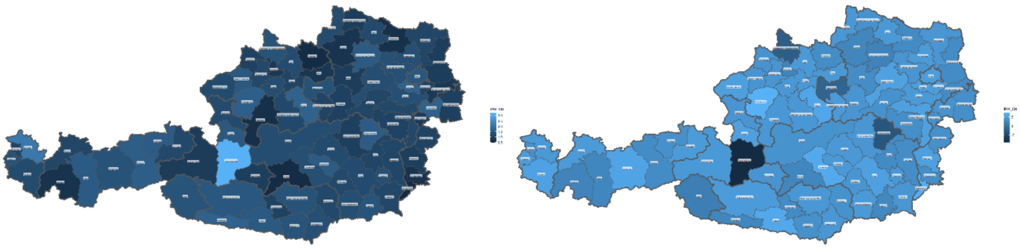


(a) 1995-2005

(b) 2005-2015

Notes: The brighter the region the higher the increase in import exposure between 1995 and 2005 or between 2005 and 2015.

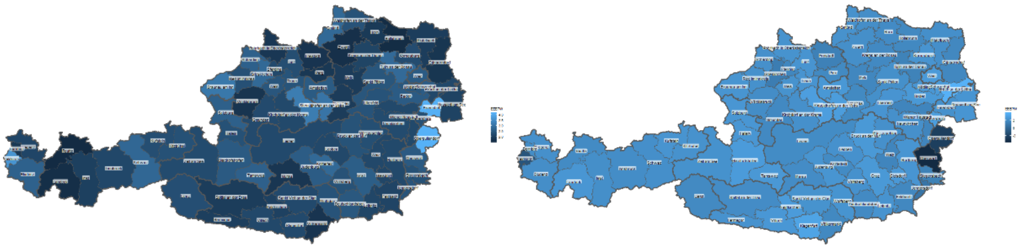
Figure A.5: Change in regionalised import competition with Eastern Europe



(a) 1995-2005 (b) 2005-2015

Notes: The brighter the region the higher the increase in import exposure between 1995 and 2005 or between 2005 and 2015

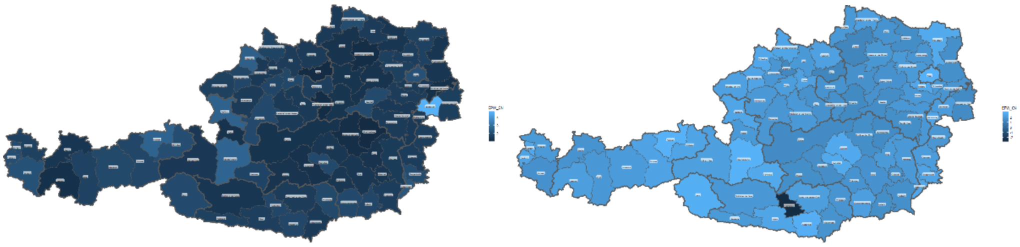
Figure A.6: Change in regionalised import competition with China in 1995-2005 (left) and 2005-2015 (right)



(a) 1995-2005 (b) 2005-2015

Notes: The brighter the region the higher the increase in export possibilities between 1995 and 2005 or between 2005 and 2015.

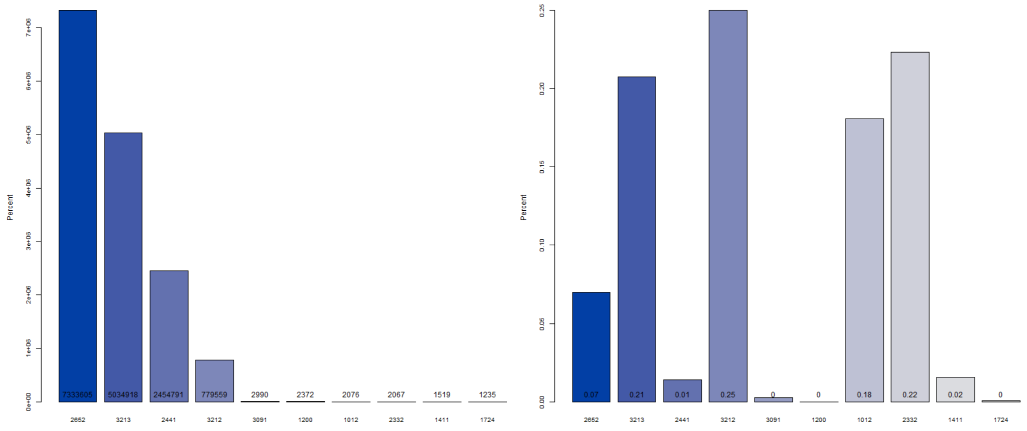
Figure A.7: Change in regionalised export competition with Eastern Europe in 1995-2005 (left) and 2005-2015 (right)



(a) 1995-2005 (b) 2005-2015

Notes: The brighter the region the higher the increase in export possibilities between 1995 and 2005 or between 2005 and 2015.

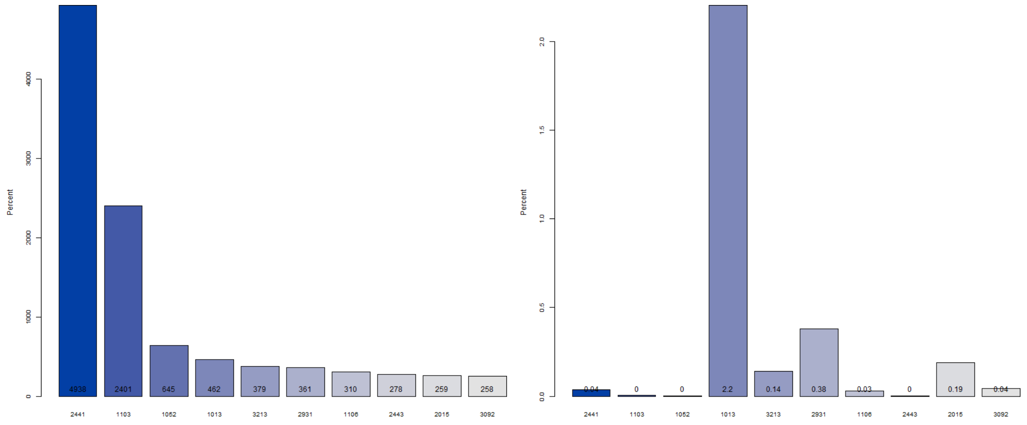
Figure A.8: Change in regionalised export competition with China in 1995-2005 (left) and 2005-2015 (right)



(a) Top10 changes in imports from Eastern Europe 1995-2005

(b) Employment shares in 1995

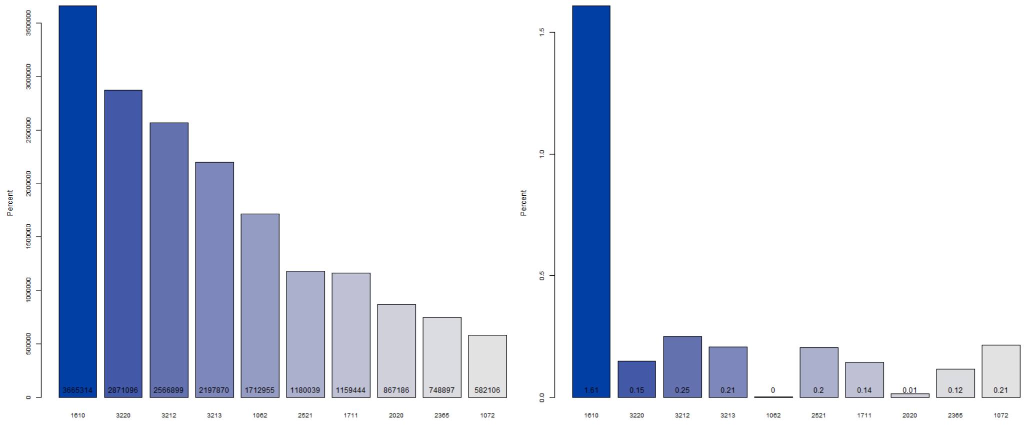
Figure A.9: Top10 import growth rates from Eastern Europe between 1995 and 2005 (left) and sectoral employment shares in 1995 (right), in %



(a) Top10 changes in imports from Eastern Europe 2005-2015

(b) Employment shares in 1995

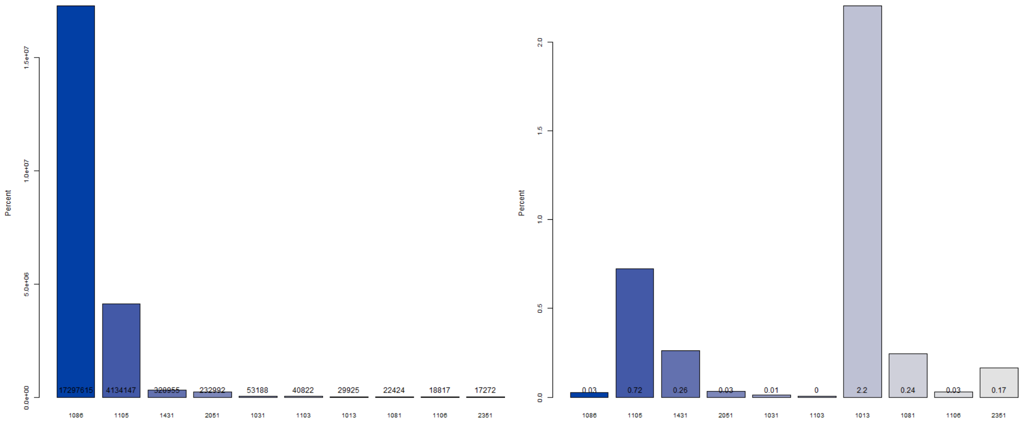
Figure A.10: Top10 import growth rates from Eastern Europe between 2005 and 2015 (left) and sectoral employment shares in 2005 (right), in %



(a) Top10 changes in imports from China 1995-2005

(b) Employment shares in 1995

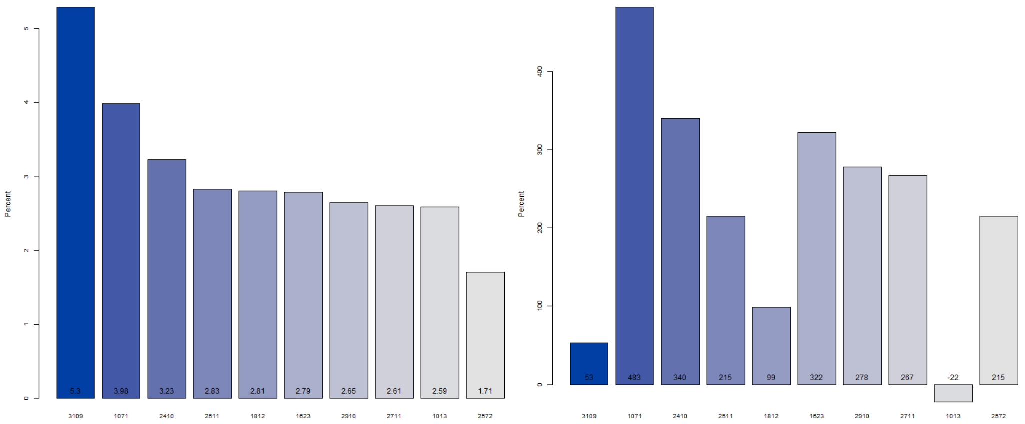
Figure A.11: Top10 import growth rates from China between 1995 and 2005 (left) and sectoral employment shares in 1995 (right), in %



(a) Top10 changes in imports from China 2005-2015

(b) Employment shares in 2005

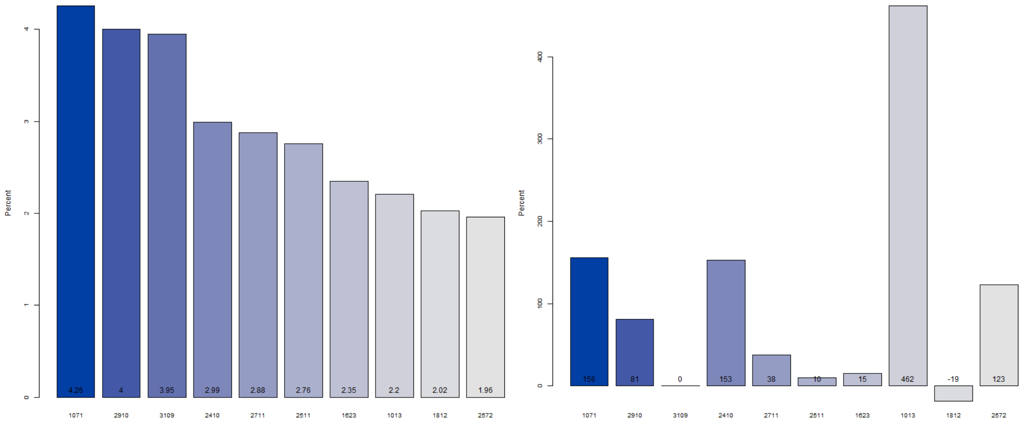
Figure A.12: Top10 import growth rates from China between 2005 and 2015 (left) and sectoral employment shares in 2005 (right), in %



(a) Top10 employment shares in 1995

(b) Changes in imports from Eastern Europe 1995-2005

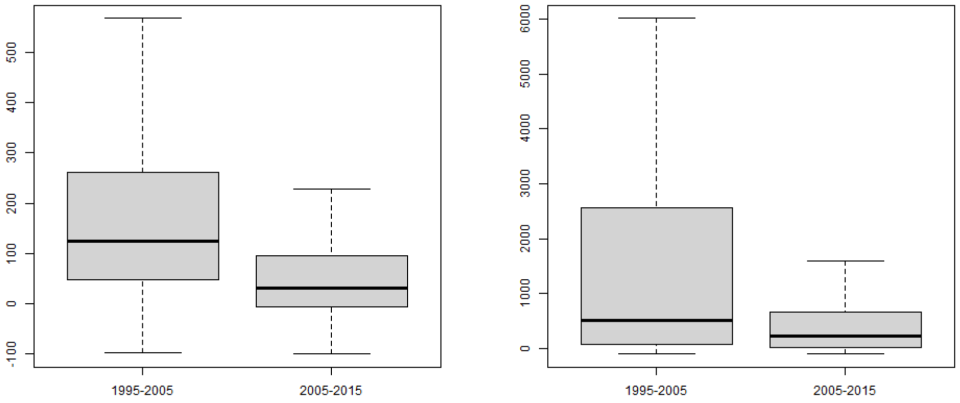
Figure A.13: Top10 largest industries in 1995 (left) and their changes in imports to Eastern Europe 1995-2005 (right), in %



(a) Top10 employment shares in 2005

(b) Changes in imports from Eastern Europe 2005-2015

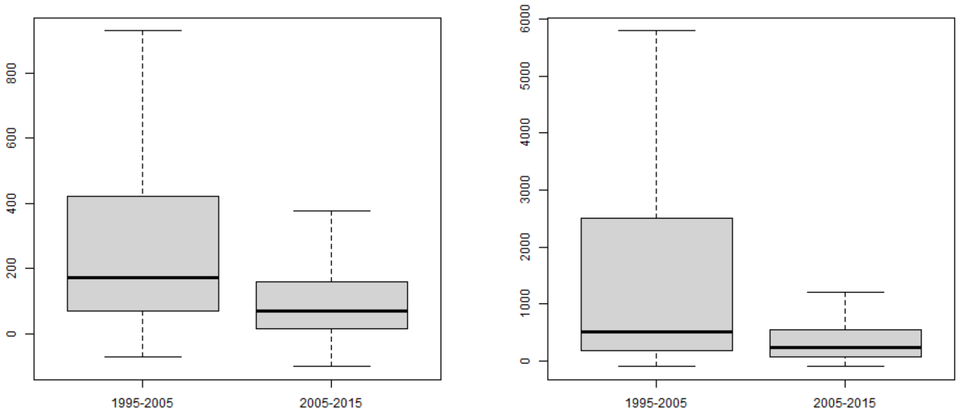
Figure A.14: Top10 largest industries in 2005 (left) and their changes in imports to Eastern Europe 2005-2015 (right), in %



(a) Eastern Europe (b) China

Notes: The whiskers are calculated as the distance of 1.5 times the interquartile range Outliers from above the upper quartile and below the lower quartile. All other observed data points outside the boundary of the whiskers are not plotted.

Figure A.15: Boxplot of percentage change in sectoral imports from Eastern Europe and China



(a) Eastern Europe (b) China

Notes: The whiskers are calculated as the distance of 1.5 times the interquartile range Outliers from above the upper quartile and below the lower quartile. All other observed data points outside the boundary of the whiskers are not plotted.

Figure A.16: Boxplot of percentage change in sectoral exports to Eastern Europe and China

REGION
The Journal of ERSA
Powered by WU

Volume 11, Number 1, 2024, 27–53
DOI: 10.18335/region.v11i1.491

journal homepage: region.ersa.org
ISSN: 2409-5370

The influence of underlying conditions of countries on the COVID-19 lethality rate

Grace Carolina Guevara Rosero[1], Eymy Coralia Illescas Navarrete[1]

[1] Escuela Politécnica Nacional, Quito, Ecuador

Received: 21 April 2023/Accepted: 10 January 2024

Abstract. The management of the COVID-19 pandemic not only depends on the stringency measures established by governments but more importantly on the underlying capacity of territories in economic, health and sanitary infrastructure. This study aims to identify how the underlying conditions of countries influence their COVID-19 lethality rate. To do so, a classification of countries is first developed by the k-means partitioning method, using COVID-19-related variables such as the lethality rate, the contagion growth rate and the number of days with respect to China. Based on the resulting groups of countries of the first stage, Tobit and Ordinary Least Squares regressions are estimated to determine the effect of the underlying characteristics of countries on their COVID-19 lethality rate. Risks factors which increase the lethality rate in countries are the contagion growth rate, the trade flow with China, the age composition of the population and, to a lesser extent, the population density. Factors that help reduce the lethality rate are the government effectiveness, the health infrastructure (hospital beds) and, to a lesser extent, the economic growth rate.

JEL classification: H50, O10

Key words: COVID-19, underlying conditions, clustering analysis, Tobit, OLS

1 Introduction

In January 2020, the outbreak of the novel coronavirus, known as COVID-19, was declared a Public Health Emergency of International Concern (PHEIC). Governments took action and established regulations to mitigate the negative effects of the pandemic related to the spread of the virus and therefore, the lethality. The measures ranged from travel to lockdown restrictions. However, the management of the pandemic not only depended on the stringency measures, but also on the pre-existing conditions at the territorial level. For instance, the health capacity of countries was an important factor to manage the pandemic. Countries with historically high spending in health would perform better with respect to countries with historically low spending. According to the World Health Organization (WHO 2022), the health expenditure per capita in high-income countries ($2,525.11) is 11 and 26 times higher than that of medium- low-income countries ($219.15) and low-income countries ($94.53). Accordingly, this study aims to determine how the underlying conditions of countries' health infrastructure, economic resources and demographic structures influenced their COVID-19 lethality rate.

The capacity and the quality of health and care services affect the impact of a pandemic disease. The virus can have a moderate effect on morbidity in countries with an effective organization of health systems and at the same time can be devastating in countries where health systems are deficient (WHO 2009). In the context of the COVID-19 pandemic, existing studies display mixed results about the effect of health expenditure on the lethality rate. Jeanne et al. (2023) indicated that the number of doctors corresponds with a higher number of infected people and deaths. This result is explained by the notion that more developed countries were affected more rapidly and the demand for health escalated in such a way that even the positive conditions of their health systems were insufficient. On the other hand, Perone (2021), despite finding that higher health expenditure increases the lethality rate, they concluded that the health system performance reduced the case fatality rate in Italy.

One of our research questions is: what is the effect of pre-existing health capacity and infrastructure and other underlying socioeconomic characteristics on the COVID-19 lethality rate of countries? Answers to this question are crucial for public policy actions because they give valuable information about critical underlying factors that increase the lethality rate in countries. These results create knowledge, especially for developing countries with limited economic resources, as to where to allocate resources to face shocks such as the COVID-19 pandemic.

To determine such an effect, it is important to consider that the role of underlying characteristics to face the pandemic will depend on the degree of the COVID-19 affectation of countries, which in turn depends on the time the pandemic was declared in each country. Therefore, a specific objective of this study is to determine groups of countries with similar characteristics related to the COVID-19 pandemic. Using clustering techniques, countries are first classified in function of COVID-19 variables such as the lethality rate, the contagion growth and the number of days that elapsed until the country registered the first case with respect to China. It allows the researchers to determine comparable groups of countries with similar COVID-19 measures. For this reason, this study provides an econometric analysis that shows comparable results across countries. To the best of our knowledge, it is the first time that comparable countries in terms of COVID-19 affectations have been analyzed. Previous analyses have built clusters of countries considering socioeconomic and demographic variables (Perone 2021, Guevara-Rosero 2022), which distorts the real COVID-19 risk. The resulting groups follow a geographical distribution in terms of velocity and affectation of the COVID-19 pandemic. We used COVID-19 information corresponding to the period of 360 days from the first registered COVID-19 confirmed case to determine the clusters. Tobit models or Multiple linear regression models were estimated on the identified clusters according to whether the dependent variable is limited or not to estimate the effects of underlying characteristics on the lethality rate associated with COVID-19.

The article follows a logical organizational structure. Section 2 describes the existing literature on the incidence of underlying conditions of countries on the dynamics of pandemics. Section 3 displays worldwide statistics about the COVID-19 pandemic. Section 4 describes the data and the methodology. Section 5 discusses the results and Section 6 provides a conclusion.

2 Literature review

There are two sets of factors that can determine the number of deaths from COVID-19. The first set of factors is related to the pandemic itself, that is, aspects that arise from the development of the disease such as the spread of the disease and government measures to prevent further spread and, in turn, deaths. Those measures ranged from travel restrictions, closure of businesses to total lockdowns. The second set of factors are related to the underlying characteristics of countries such as demographics, economics, health capacity and governmental effectiveness.

Regarding the variables related to the pandemic itself, the Pan American Health Organization (PAHO 2022) mentioned that the transmission rate of the virus identifies the severity of the disease and its influence on the lethality of the population. Several studies have considered this factor to explain the lethality rate. Peralta et al. (2020) concludes that a high rate of contagion by COVID-19 implies that many infected people simultaneously cause the collapse of the health system and make it difficult for seriously ill cases to access to it. Saturation of the health system was analyzed by Perone (2021) through the ratio prevalence/ ordinary beds, which explained 86% of the case fatality rate in Italy. Another factor associated with the pandemic was the stringency measures implemented by governments to curb the transmission of the coronavirus SARS-CoV-2 and reduce mortality. Chaudhry et al. (2020) and Chisadza et al. (2021) highlight that less stringent measures increase the number of deaths from COVID-19. Jinjarak et al. (2020) concludes that strict policies to curb the spread of the virus were associated with lower mortality growth rates. Sorci et al. (2020) deduces that a higher case fatality rate is reached for intermediate values of the stringency index.

Pertaining to the underlying characteristics of countries, an important characteristic that influences their lethality level is the age composition of the population. This is due to the risk of hospitalization or death from COVID-19 increasing for people older than 60 years old by reason of the existence of more factors that make them prone to severe illness (Centers for Disease Control and Prevention 2021). For Promislow, Anderson (2020) and King et al. (2020), the health risks related to the virus increase with age. Regarding the fatality rate, Rubino et al. (2020) performs a comparative analysis for the first weeks of the pandemic and shows the Italian case fatality rate is 10.6% due to the fact it is a country with a large elderly population. The average age of death from COVID-19 was 81 years in Italy.

Generally, residing in a densely populated area is a contagion risk factor for SARS-CoV-2 (de Lusignan et al. 2020) as the physical proximity of infected people in urban centers facilitates the transmission (Waltenburg et al. 2020, Rocklöv, Sjödin 2021). Likewise, Ilardi et al. (2021) identifies a significant positive linear relationship between population density and case fatality rate. Some studies highlight a negative or a non-significant relationship between population density and deaths from COVID-19, indicating that population density is not a risk factor for lethality. Places with higher density are more likely to have considerable resources to respond to the pandemic and reduce the number of deaths (Fang, Wahba 2020).

The medical equipment and staff are also factors that explain the number of deaths from COVID-19 (Ilardi et al. 2021). A low number of hospital beds causes the collapse of the health system and therefore increases deaths (Acosta 2020, Park, Cha 2020). Chaudhry et al. (2020) and Asfahan et al. (2020) obtain that the number of doctors per 10,000 inhabitants has an inverse relationship with the fatality rate. Given this, Khan et al. (2020) suggests that building an effective multidimensional healthcare capacity is the means to mitigate deaths from future cases. Although a large number of doctors assumingly decreases the lethality rate, Jeanne et al. (2023) demonstrate a larger number of doctors was associated with a higher number of infected cases and deaths at the beginning of the pandemic. This indicates that in the case of European countries, the positive healthcare system was not sufficient to face the pandemic waves.

Regarding the economic level of countries, measures such as GDP or health spending are used in the literature. Asfahan et al. (2020) used a univariate regression to find that GDP per capita is negatively correlated with the case fatality rate from COVID-19. In the study by Chaudhry et al. (2020), countries with a higher GDP per capita recorded a higher number of deaths per million inhabitants. This reflects more widespread testing in those countries and greater transparency in reporting, as well as increased accessibility to air travel and international vacations in wealthier countries. It is worth mentioning that, during the first months, the burden of the pandemic was mainly focused on high and middle-income countries in Asia, Europe and North America (Chisadza et al. 2021) as they were most connected to China (Jeanne et al. 2023). In other words, countries with better conditions suffer a greater impact from the virus (Zevallos, Lescano 2020). Although the pandemic quickly reached countries with a high economic level, the consequences were more severe for less developed countries. The report presented by the Imperial University of London on the evaluation of the impacts of the pandemic on disadvantaged and vulnerable populations highlighted that the risk of death from

COVID-19 increases with poverty (Winskill et al. 2020). As for health spending, Khan et al. (2020) conclude that this variable did not reach statistical significance but was positively associated with fatal cases at the global level. For the Italian case, Perone (2021) deduced that the average public health expenditure per capita increases the death rate in Italian regions. While unexpected, the author explains that this variable is only one dimension of health system performance that globally was significant and negative for the death rate.

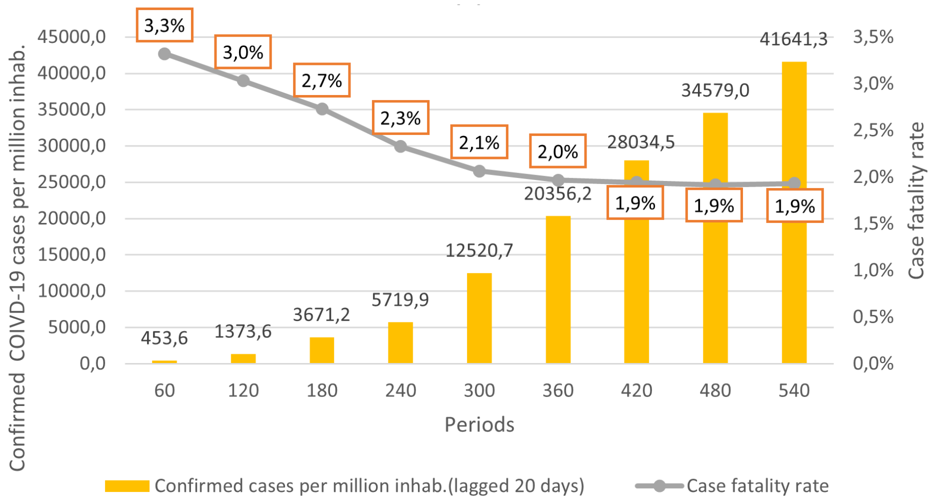
Trade flows and human capital mobility between countries are vital to understand the pandemic. Nations are strongly interconnected as a result of globalization, raising the potential of the pandemic spread (Opertti, Mesquita-Moreira 2020, Spyrou et al. 2016). Globalization contributed to the importation of cases (Abellán et al. 2020). However, just as global value chains were the main transmission channel for the first countries infected by the disease, they were also the main transmission channel for the effects of COVID-19 on world trade (Comisión Económica para América Latina y el Caribe; Naciones Unidas 2020).

Aside from inadequate health infrastructure, poor governance significantly complicates outbreak preparedness and response. It limits the state's ability to act effectively, which has devastating consequences for loss of human life, economic destabilization and social chaos (GPMB 2019). Liang et al. (2020) demonstrates that for a short-term crisis like the COVID-19 outbreak, government effectiveness is critical to respond efficiently and ensure effective policies that reduce case fatality rates (Serikbayeva et al. 2021). By contrast, Toshkov et al. (2021) conclude, based on robust models, that highly perceived capacity for government effectiveness provides false confidence that results in higher infections and deaths.

3 Pandemic Statistics

There are examples among positive COVID-19 cases in which the illness follows its course and ends with recovery, but there are other cases which get worse, requiring hospitalization, potentially resulting in death. A strategy to contain the spread of COVID-19 that had been used by many governments was the lockdown. This approach restricted face-to-face contacts among people. Once stringency measures were in place, people reduced contact with others (five days are considered in this study). Once a person became infected, the illness lasted approximately 15 days, but if it worsened, it could last longer until death (Javed 2020) (20 days are considered in this study). On this basis, the statistics illustrated in Figure 1 highlight the number of COVID-19 infections per million inhabitants at day t-20 and the case fatality rate at day t. The graph is specified by periods, namely, 60, 120, 180, 240, 300, 360, 420, 480 and 540 days from the first registered COVID-19 case in each country. The numbers are cumulative. On average, the worldwide cumulative fatality rate at the 60th day was 3.3% and decreased with time until reaching 2% at the 360th day. From this period to 540 days, the fatality rate had been stable at 1.9%. By contrast, the cumulative number of infection cases per million inhabitants had been increasing from 453.6 confirmed cases at the beginning of the pandemic to 41,641 infected people per million inhabitants at 540 days.

These COVID-19 statistics vary between countries which were rapidly infected (less than 75 days with respect to the first COVID-19 case registered in China) and slowly infected countries (more than 75 days with respect to the first COVID-19 case registered in China). The period of 75 days is the average number of days that elapsed between the first COVID-19 case in China and the first registered COVID-19 case in other countries. 146 countries registered the first COVID-19 case before the average of 75 days (hereafter, named as rapidly infected countries) and 67 countries that registered the first COVID-19 case after the average of 75 days (hereafter, named as slowly infected countries). Figure 2 shows a significant difference of COVID-19 confirmed cases per million inhabitants between rapidly and slowly infected countries in all periods. At the beginning of the COVID-19 pandemic (60-day period), the number of COVID-19 cases in rapidly infected countries was on average 535 per million inhabitants meanwhile it was 280 confirmed cases per million inhabitants in slowly infected countries. Rapidly infected countries



Source: European Center for Disease Prevention and Control (ECDC)
Notes: Own elaboration (March 2023)

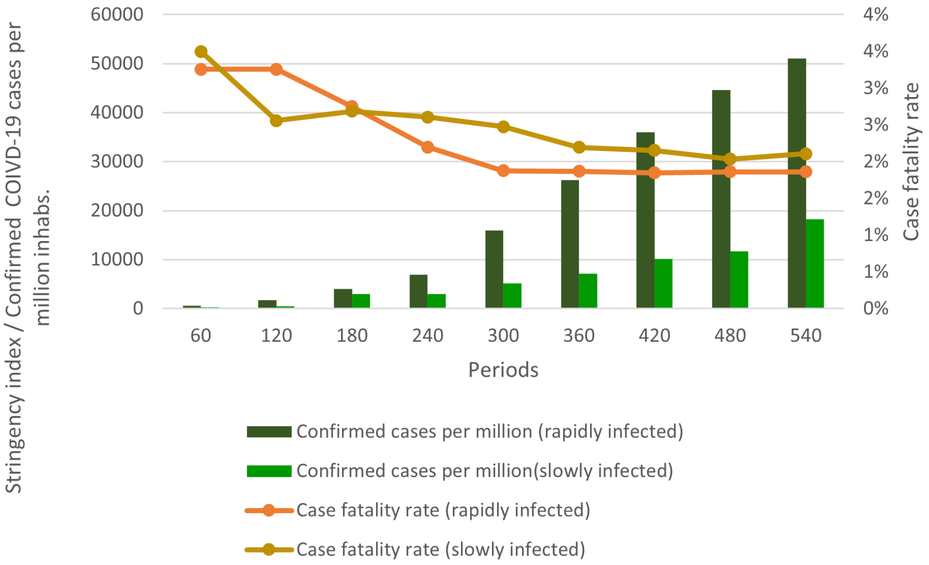
Figure 1: COVID-19 fatality rate, confirmed cases and the stringency index by periods

with the highest value of cumulative number of COVID-19 cases per million inhabitants at the 60th day were Vatican, San Marino, Andorra, Luxembourg and Iceland. Those with the lowest value for rapidly infected countries were Russia, Cambodia, Sri Lanka, India and Nepal. Slowly infected countries with the highest value of cumulative number of COVID-19 cases per million inhabitants at the 60th day were the Falkland Islands, Isle of Man, Montserrat, Bermuda and Sao Tome and Principe and those with the lowest value for slowly infected countries were Nicaragua, Uganda, Burundi, Papua New Guinea and Angola. The cumulative number of COVID-19 cases per million inhabitants at the 540th day was 51,097 in rapidly infected countries whereas it was only 18,241 in slowly infected countries. Rapidly infected countries with the highest value of cumulative number of COVID-19 cases per million inhabitants at the 540th day were Seychelles, Andorra, Czechia, Bahrain, Gibraltar and San Marino and those with the lowest number for rapidly infected countries were Macao, Vietnam, New Zealand and the Democratic Republic of Congo. Slowly infected countries with the highest value of cumulative number of COVID-19 cases per million inhabitants at the 540th day were Montenegro, British Virgin Islands, Isle of Man, the Bonaire, Sint Eustatius and Saba islands (BES) and the Turks and Caicos Islands. Those with the lowest number were Tanzania, Niger, Yemen, Chad and New Caledonia. However, slowly infected countries recorded a higher lethality rate than rapidly infected countries, except for the period of 120 days. Therefore, countries with fewer days of preparation had a lower case-fatality rate than those with more time to face the health crisis This can be explained by a higher level of development of rapidly infected countries with respect to slowly infected countries. A decreasing trend in the fatality rate from COVID-19 after 180 days of the pandemic is observed in both cases.

3.1 Pandemic statistics per period

Figure 3 shows the number of COVID-19 confirmed cases by periods. Countries that became infected later, with respect to China, report a lower number of COVID-19 cases for all periods of time, except for the 61–180-day period. The number of COVID-19 cases in rapidly infected countries increased emphatically from the periods of 61–180 days to 181-360 days and remains around 25000 cases per million inhabitants for the 361-540 day period. Whereas the number of COVID-19 cases per million inhabitants in slowly infected countries increased over time.

The case-fatality rate shown in Figure 4 indicates that the lethality rate starts at high levels (3.3%) and decreases to 1.8% after one and a half years since the beginning of



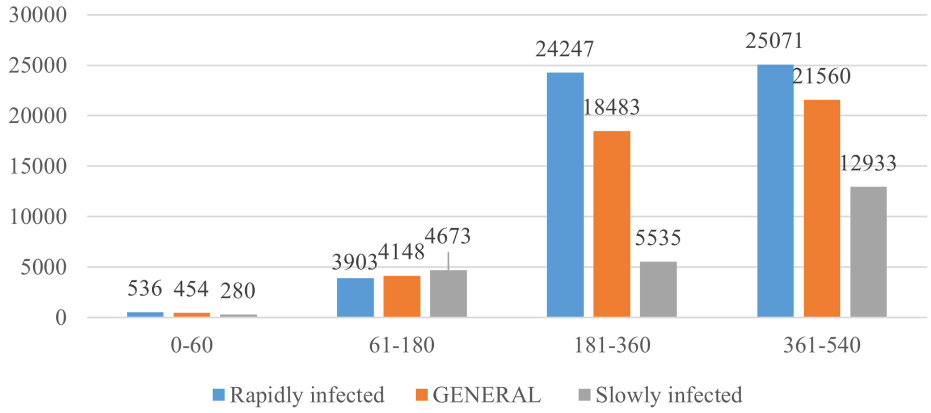
Source: European Center for Disease Prevention and Control (ECDC)
Notes: Own elaboration (March 2023)

Figure 2: COVID-19 fatality rate, confirmed cases and the stringency index by periods and types of countries

the pandemic, recording a reduction of 45%. This reduction of the lethality rate can be explained by the rollout of vaccines in 2021. Although the COVID-19 pandemic arrived later to certain countries (slowly infected countries), they recorded a higher case fatality rate (3.5%) during all periods of the pandemic except for the period of 61 to 180 days.,

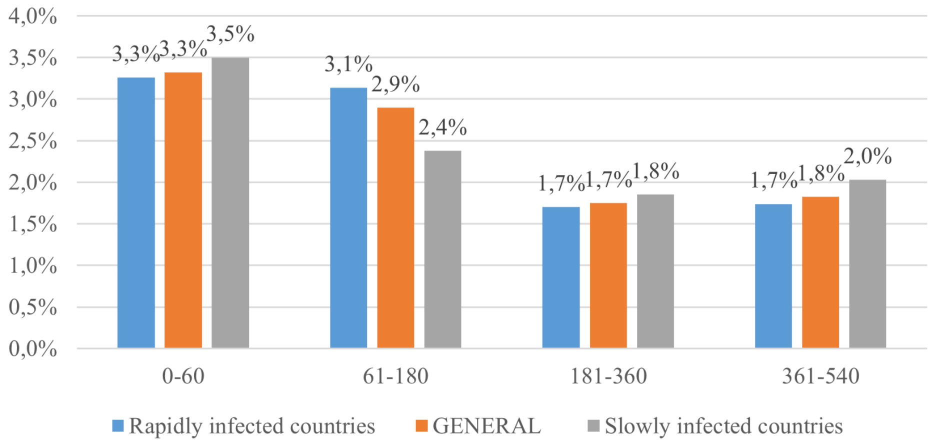
Table 1 displays the average number of COVID-19 confirmed cases per million inhabitants, the growth rate of contagion, the fatality rate and the number of days with respect to China for each continent for different periods. The first two variables are analyzed by periods (0 to 60 days; 61 to 180 days;181 days to 360 days;361 days to 540 days), while accumulated data is presented for the variables case fatality rate and days with respect to China.

The regions that registered the first COVID-19 confirmed case very rapidly with respect to China were North America (50 days), Asia (54 days) and Europe (57 days). They had less than two months to prepare for the pandemic. Africa and Latin America were the last regions to register the first COVID-19 case (78 days and 74 days, respec-



Source: European Center for Disease Prevention and Control (ECDC)
Notes: Own elaboration (March 2023)

Figure 3: COVID-19 cases per million inhabitants per periods



Source: European Center for Disease Prevention and Control (ECDC)
Notes: Own elaboration (March 2023)

Figure 4: Case fatality rate by COVID-19 per periods

tively)[1]. Regarding the number of confirmed cases per million inhabitants, Europe was the region that registered the highest number of confirmed cases for the 0–60-day period, roughly four times that of North America, the region with the second most infections. The number of infected people grew drastically between periods in all regions. The average growth rate between all periods was 341.8% for Africa, 441.45% for LAC, 790.6% for Oceania, 459.4% for Asia, 518.7% for Europe and 337.4% for North America. However, the growth rate of the number of infected people between the 180–360-day period and 360–540-day period is lower than the growth rate between previous periods. The growth rate between the last periods was 153.2% for Africa, 94.4% for LAC, 199% for Oceania, 69.9% for Asia, -33.6% for Europe and 15.56% for North America[2].

Regarding the average of daily growth rates of confirmed cases within periods (column 2), Europe recorded the highest daily growth rate in the first period, followed by Asia. Although the SARS-CoV-2 virus appeared to/in Africa last, this region recorded a higher daily growth rate than Latin America. The daily growth rate of infected people decreased across periods for all regions. Although the number of confirmed cases grew as time progressed, the contagion rate decreased. Most regions (except Oceania) registered contagion rates greater than 10% for the period from 0 to 60 days. The figure declined so that the contagion growth rate did not exceed 1.05% for the period of 361 to 540 days.

The most affected region for fatality rate in most periods is Latin America and the Caribbean (LAC). During the first two months, the region registered the highest fatality rate (5.40%) despite the contagion level and growth rate not being the highest. The regions of Europe and North America registered the highest fatality rate on average for the 180-day period. However, by 360 days and 540 days, Africa recorded the highest lethality rate despite the low rate of spread. This could be explained by their development level (Winskill et al. 2020, Sanmartín-Durango et al. 2019).

4 Data and methodology

The European Center for Disease Prevention and Control (ECDC) and the Blavatnik School of Government in Oxford are the main data sources used in this study. Data regarding the underlying conditions of countries in terms of health infrastructure, demographics and economics were sourced from the World Bank (WB), the World Health Organization (WHO) and the Global Innovation Index (GII). Across-sectional database was obtained with this information. Three types of variables were considered: i. COVID-19

[1] It is worth noting that Latin America and Africa are regions where testing was not enough so there could be an under-registration of cases.

[2] Growth rate of the number of infected people between periods is calculated as $\frac{\text{final value}}{\text{initial value}} - 1$.

Table 1: COVID-19 statistics by continent

Continent	COVID-19 confirmed cases per million	Daily growth rate of confirmed cases	Case fatality rate	Days respect to China
	0-60 days		60 days	
Africa	155.78	13.34%	3.28%	77.56
Latin America & Caribbean (LAC)	467.34	11.94%	5.40%	74.07
Oceania	115.97	8.55%	0.29%	66
Asia	351.18	15.91%	2.13%	54.33
Europa	2165.73	17.99%	4.09%	56.96
North America	599.53	12.6%	2.68%	50
	0-180 days		180 days	
Africa	1192.03	2.29%	2.26%	77.56
Latin America & Caribbean	5431.44	2.40%	2.47%	74.07
Oceania	529.81	1.46%	1.62%	66
Asia	4327.32	2.30%	2.15%	54.33
Europa	2949.19	1.09%	4.20%	56.96
North America	3782.37	1.93%	4.22%	50
	0-360 days		360 days	
Africa	3659.7	0.67%	2.13%	77.56
Latin America & Caribbean	14545.09	0.86%	2.13%	74.07
Oceania	10627.81	0.74%	1.53%	66
Asia	11940.44	0.86%	1.80%	54.33
Europa	48761.99	1.41%	1.90%	56.96
North America	21397.86	0.84%	1.50%	50
	0-540 days		540 days	
Africa	9265.57	0.43%	2.18%	77.56
Latin America & Caribbean	28269.13	0.69%	2.16%	74.07
Oceania	22208.66	1.04%	1.24%	66
Asia	20291.41	0.77%	1.92%	54.33
Europa	32399.34	0.29%	1.74%	56.96
North America	24727.97	0.72%	1.17%	50

Notes: Elaborated by the authors

related variables such as the contagion growth rate and the stringency index; ii. structural variables such as the Government Effectiveness Index, GDP per capita growth, health expenditure as a percentage of GDP; iii. conjunctural variables such as the population aged more than 65 years, population density, hospital beds, doctors and trade flows with China. An average over the last five available years is used for the former. For the latter, the data of the last available year is used. We considered the day at which the first COVID-19 confirmed case was registered in each country by reason that the SARS-CoV-2 virus arrived in each country at different times. Therefore, the cross-country analysis is comparable. The database consists of 211 countries for the clustering analysis and 128 countries for the econometric analysis due to the limited availability of some variables for some countries.

4.1 Method

This study was conducted in two phases. In the first phase, clusters of countries based on COVID-19 variables were determined. The clusters of countries with similar characteristics related to the COVID-19 pandemic were used for estimations in the second phase. This allowed determining the effect of underlying socioeconomic characteristics on countries with similar affectations derived from the pandemic. Tobit and Ordinary Least Squares regressions were estimated. The choice of the Tobit models corresponded to the existence of a limited dependent variable (VDL). The lethality rate in many countries was zero at the beginning of the COVID-19 pandemic. Therefore, the dependent variable was limited since it maintained the zero-limit, and some observations hit this limit. The censored sample is representative of our group of countries and since many values record zero, the mean tends to be low. The OLS selection is derived from the non-existent zero lethality rates cases in some groups of countries.

4.1.1 First phase: Clustering analysis

To determine the clusters of countries in relation to their evolution in terms of the pandemic, the k-means method clustering technique was employed. The k-means method was developed by MacQueen (1967) and Lloyd (1982). It partitions the database of n objects in k clusters, such that the sum of square distances (see equation 1) between the observation, p, and the centroid of the cluster, c_i, is minimized.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_1)^2 \tag{1}$$

The k-means algorithm first partitions objects into k nonempty subsets. It then calculates the clusters' centroids (mean point) of the current partitioning. Third, it assigns each object to the cluster with the nearest centroid. And last, it stops when the assignment is stable so that clusters do not change (Han et al. 2022, p. 451).

We used a set of 211 countries to determine the clusters. The data for 360 days was then applied for the variables: lethality rate, contagion growth rate and number of days with respect to China. The number of clusters was chosen based on the Elbow method that displays the intra-class variance according to the number of clusters. The optimal number of clusters is identified given a threshold at which the intra-class variance does not significantly decrease. The study employed the Elbow method, which indicated that the intra-class variance did not significantly decrease after three clusters.

4.1.2 Second phase: Estimation models

Models were estimated for each cluster (group A and group B[3]) and each period (60 days, 180 days, 360 days and 540 days). Tobit models and OLS models were estimated. The Tobit Model was used when the dependent variable, y, was zero for a non-trivial fraction of the population and when the OLS predictions were negative. This model had an approximately continuous distribution through positive values (Gujarati, Porter 2010, Woolridge 2010). The Tobit model for corner solution responses is estimated by Maximum Likelihood Method. This model involves non-negative predicted values that have sensible partial effects on the range of independent variables. The observed response, y, is expressed in terms of an underlying latent variable as shown in equation (2).

$$\begin{aligned} y^* &= \beta_0 + \mathbf{X}\boldsymbol{\beta} + u \qquad (2) \\ y &= \max(0, y^*) \\ u|x &\sim N(0, \sigma^2) \end{aligned}$$

where y^* satisfies the assumptions of the classic linear model and has a normal and homoscedastic distribution with a linear conditional mean. In our case, y_i is the lethality rate, which is calculated as the number of deaths from the disease divided by the total number of cases in a specific period (Moreno et al. 2020). This measure represents the severity of which COVID-19 affected the population. X is a vector of explanatory variables, which are described in Table 2.

A multiple linear regression model was used for groups of countries that did not have a limited dependent variable. The specification is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i \tag{3}$$

where Y_i is the COVID-19 lethality rate as in the Tobit model, ϵ_i is the random error term.

Descriptive statistics of variables concerning the underlying conditions of countries and COVID-19 related variables are displayed in Table 3. According to these statistics,

[3] Group C has insufficient observations (7) to run a model.

Table 2: Description of independent variables and expected results

Independent variable	Description	Exp. sign	Supporting literature
Contagion growth rate at t-20 days	Average of the daily growth rates at a certain period (60 days, 180 days, 360 days and 540 days). The daily growth rate is calculated as $\frac{\text{cases}_i - \text{cases}_{i-1}}{\text{cases}_{i-1}}$ and lagged by 25 days.	(+)	Peralta et al. (2020)
Stringency index at t-25 days	A 0-100 scale index (100 = strictest), based on nine response indicators including school closures, workplace closures, and travel bans. This variable is lagged in 25 days.	(-)	Jinjarak et al. (2020), Chisadza et al. (2021)
Government effectiveness index	The index reflects the population perception of the quality of public services, the quality of the civil service, the degree of governmental independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. This index is averaged for the last five available years.	(-)	Liang et al. (2020), Toshkov et al. (2021)
Population density	Calculated by dividing the total population of a country i by the surface in square kilometers, for the last available year	(ambiguous)	Chaudhry et al. (2020)
Percentage of the population older than 65 years old	Percentage of the population older than 65 years, for the last available year	(-)	de Lusignan et al. (2020)
GDP growth rate	Average of the GDP growth rate for the last 5 years	(-)	Asfahan et al. (2020), Chaudhry et al. (2020)
Health spending as a percentage of GDP	Average percentage of GDP allocated to Health Expenditure for the last 5 years	(-) / no significant	Asfahan et al. (2020), Khan et al. (2020)
Exports from China	Exports between China and country i in thousands of dollars for the last available year, in logarithm.	(+)	Abellán et al. (2020)
Hospital beds per 10000 inhabitants	Number of hospital beds per 10,000 inhabitants, for the last available year.	(-)	Park, Cha (2020), Acosta (2020)
Number of doctors per 10000 inhabitants	Number of doctors per 10,000 inhabitants, for the last available year	(-)	Chaudhry et al. (2020), Asfahan et al. (2020)

the contagion growth rate decreases over time. During the first period (60 days), the daily average contagion growth rate is 18%. It decreases to 2.6% during the last period (540 days). The high contagion rate at the beginning was caused by the ignorance about the forms of contagion and the dynamics of the virus spread. For instance, at the beginning, recommendations suggest the use of face masks only for ill people. The average stringency index reduced over time. The average of the government effectiveness is 47.45 points out of a maximum of 100 points. There is a high dispersion of this variable, indicating high heterogeneity between countries. the standard deviation (1507.008) of population density is greater than the mean (332.013). The average percentage of population older than 65 years old is 9.20%. The average number of doctors per 10,000 inhabitants is 21.19 and the average number of hospital beds per 10,000 inhabitants is 27.6. There is high dispersion in these measures, reflecting high disparity across countries. The average percentage of the population that has access to drinking water is 87.53%. On average, the annual GDP growth rate for the last five years was 2.868%. However, the dispersion is high since it ranges from -10.793% to 10.076%. The percentage of GDP allocated to health spending is 2.868%, and the standard deviation is low. China's exports to each country are on average 1.2 billion dollars.

Table 3: Descriptive statistics independent variables

Variable	Obs.	Min.	Max.	Std. Dev.	Mean
Lagged contagion growth rate at 60 days $(t-20)$	215	0	0.65	0.104	0.18
Lagged contagion growth rate at 180 days $(t-20)$	214	0	0.1879	0.028	0.065
Lagged contagion growth rate at 360 days $(t-20)$	211	0	0.091	0.013	0.036
Lagged contagion growth rate at 540 days $(t-20)$	205	0.001	0.061	0.0076	0.026
Lagged stringency index at 60 days $(t-25)$	181	0	100	25.87	70.91
Lagged stringency index at 180 days $(t-25)$	180	11.11	94.44	19.44	57.42
Stringency index at 360 days $(t-25)$	179	2.78	90.74	19.99	57.03
Stringency index at 540 days $(t-25)$	177	2.78	93.52	17.6	51.46
Government effectiveness	141	0.00	99.583	23.356	47.447
Percentage of population older than 65 years old	182	1.157	28.002	6.496	9.1965
Population density	196	0.137	19196	1507.008	332.013
GDP growth (average of the 5 last years)	190	-10.793	10.076	2.731	2.868
Current health expenditure (average of the 5 last years)	177	1.725	17.41	2.563	6.447
Hospital beds per 10000 inhabitants	172	1	129.8	22.875	27.6
Number of doctors per 10000 inhabitants	161	0.23	82.95	18.912	21.193
Logarithm of Chinese exports to each country i in thousands of dollars for the last available year	211	6.03	418584249.5	3.96e+07	1.20e+07

Source: Data from the World Bank, World Health Organization, Global Innovation Index, the Blavatnik School of Government in Oxford and the European Center for Disease Prevention and Control.
Notes: Own elaboration

The assumptions of the classic Gauss model were tested for robustness and valid interpretation of the estimations (Gujarati, Porter 2010, p. 61), and their results are shown in Table 4. According to the homoscedasticity test (White), the null hypothesis, that variance of errors is constant, was not rejected for all models. According to the Jarque-Bera test, all models presented normal residuals, except for the model of the 180-day period. For such a model, the normality was corrected by eliminating outlier observations (8 observations were eliminated). The Ramsey test results indicate there is no problem of omitted variable bias for most of the models, except for the models at 180 days. However, since the rest of the models have the same variables, the omitted variable bias in the 180-day models does not constitute a problem. According to the VIF (which is lower than 10), there is no multicollinearity in any model.

5 Results

5.1 *Cluster analysis*

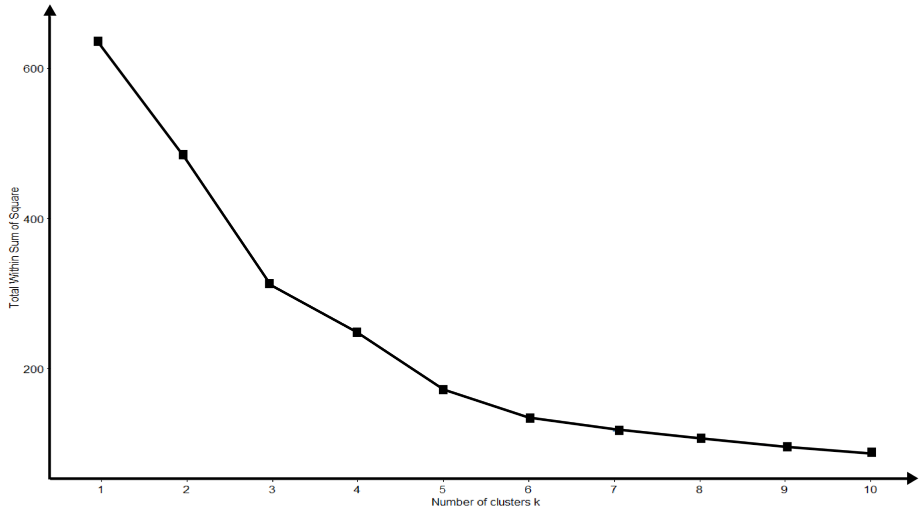
COVID-19 related variables were used for the clustering analysis of countries, namely the COVID-19 lethality rate, the contagion growth rate at t-20 and the number of days that elapsed before the first confirmed case with respect to China. The clustering analysis was based on the 360-day period (360 days after the first confirmed case in each country). The Elbow method represented in Figure 5 indicates the total within sum of square as a function of the number of clusters. The optimal number of clusters is three because this is when the total within sum of squares begins to level off. This is the number of clusters that allowed for similar observations within clusters and different observations between clusters.

Figure 6 and Table 3 displays the resulting clusters using the *k*-means partition method for the 360-day period. This indicates the standardized mean for each variable by group. Clusters are distributed geographically (shown in Figure 7).

From these results, three clusters were obtained:

Table 4: Model validation

MODEL	VIF	White Test	Normality Test	Ramsey test	Method
		GROUP A			
60 days	2.35	0.2557	0.6116	0.5208	TOBIT
180 days	2.37 → 2.32	0.4210 → 0.2675	0.0294 → 0.0504	0.0183	TOBIT
360 days	2.34	0.4417	0.2696	0.0640	TOBIT
540 days	2.37	0.4596	0.2399	0.0616	TOBIT
		GROUP B			
60 days	2.40	0.4334	0.3136	0.0634	TOBIT
180 days	2.28	0.4334	0.3874	0.0225	TOBIT
360 days	2.26	0.4334	0.1343	0.1008	OLS
540 days	2.21	0.4334	0.5524	0.2359	OLS
		GENERAL MODEL			
60 days	2.26	0.0716	0.3370	0.0503	TOBIT
180 days	2.30 → 2.28	0.5684 → 0.3225	0.0033 → 0.1160	0.0000	TOBIT
360 days	2.31	0.8641	0.2780	0.5819	TOBIT
540 days	2.25	0.9169	0.2987	0.4612	TOBIT



Notes: Own elaboration

Figure 5: Optimal cluster number using the Elbow Method

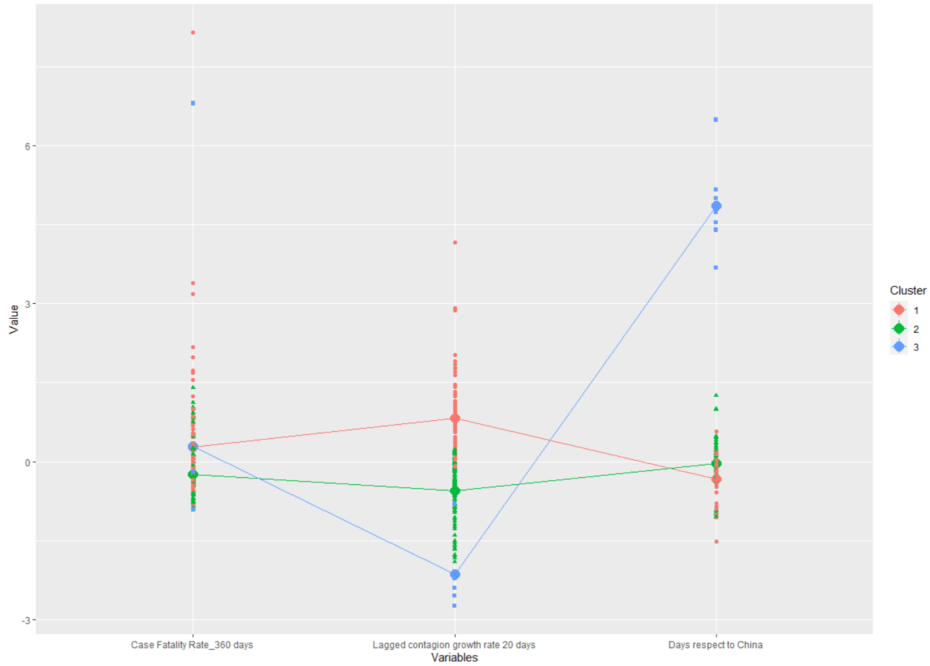
CLUSTER A: Rapidly infected countries with high lethality rate and high contagion growth rate

CLUSTER B: Less rapidly infected countries with very low lethality rate and moderate contagion growth rate

CLUSTER C: Slowly infected countries with very high lethality rate and very low contagion growth rate

Countries of Cluster A are mainly located in Europe, Asia and the Americas, except for the Caribbean. Countries of Cluster B are mainly located in Africa, Oceania and the Caribbean. Countries of Cluster C are mainly distant islands.

Cluster A encompasses 93 countries (see Appendix A), distributed around the world: Europe (36), Asia (24), Africa (15), North America (3, including Mexico), South America (9), Central America and the Caribbean (6). On average, these countries registered COVID-19 confirmed cases 58 days after the first case in China. They had less than two months to prepare themselves to face the pandemic. They register the highest contagion growth rate (4.02%) for this reason. The rapid arrival of the pandemic to these countries can be explained by their proximity to China in geographical and trade



Notes: Own elaborations

Figure 6: Graphic description of characteristics of clusters for 360 days

Table 5: Standardized means of COVID-19 variables by clusters

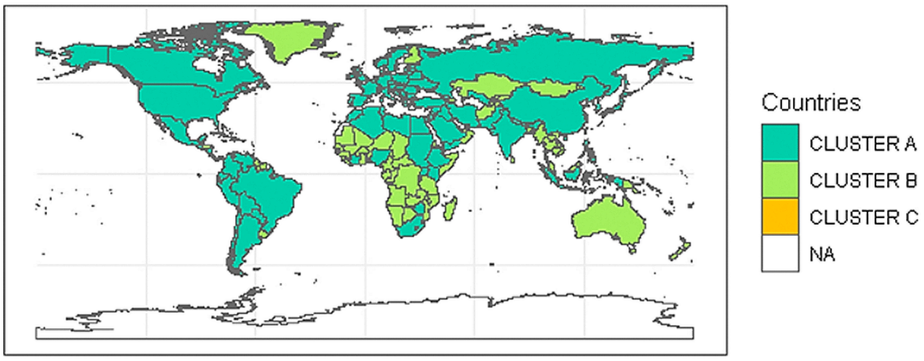
Cluster	Case Fatality Rate	Lagged contagion growth rate	Days with respect to China
Cluster 1 (A)	0,27	0,82	-0,33
Cluster 2 (B)	-0,25	-0,55	-0,03
Cluster 3 (C)	0,29	-2,14	4,85

terms. Trade openness and human mobility were important drivers of the COVID-19 worldwide spread. As a result of globalization, all nations are strongly interconnected, which worsened the spread of COVID-19 (Opertti, Mesquita-Moreira 2020, Spyrou et al. 2016). Another characteristic of countries in cluster A is the high lethality rate (2.5%). Cluster A encompasses countries from all continents. This indicates that COVID-19 dynamics tend to be similar across countries after a year following the first confirmed case in each country. Different dynamics are observed across countries when analyzing only a month after the first confirmed case was registered in each country, as revealed by Guevara-Rosero (2022).

Cluster B encompasses 111 countries (see Appendix A) from Africa (39), South America (4: Falkland Islands, Guyana, Suriname & Uruguay), Central America and the Caribbean (22), Asia (22), Europe (13), North America (2) and Oceania (6). On average, the SARS-CoV-2 virus arrived 71 days after the first registered case in China. The average lethality rate was 1.04%, whereas the contagion growth rate was 2.89%. Most countries in this group are not very interconnected with China in terms of human mobility or trade.

Cluster C encompasses only seven islands: in Oceania (Micronesia, Marshall Islands, Solomon Islands, Vanuatu, Wallis and Futuna & Samoa) and in Africa (Saint Helena). Due to their distant geographical position and low accessibility, these were the last countries that registered COVID-19 cases, on average, 308 days after China. They had more than 10 months to prepare themselves.

It is worth noting that the resulting groups do not follow a geographical pattern by continents; they are clustered by function of underlying mechanisms such as inter-



Notes: Own elaboration

Figure 7: Geographical distribution of clusters

connectedness with China and their response to the pandemic. To better understand the distinctions between clusters, socioeconomic and demographic characteristics were analyzed. Appendix B identifies those countries of Cluster A recorded a higher average government effectiveness index (49.7) than countries in Cluster B (44.7). In terms of demographic characteristics, countries in Cluster A recorded a higher percentage of population older than 65 years (11.4%), compared to countries in Cluster B (6.8%) and in Cluster C (4%). The average of the last five years of GDP growth was quite similar across clusters (2.8% in Cluster A; 2.9% in Cluster B and 3.03% in Cluster C). Regarding the health sector, countries in Cluster A recorded a higher average of the number of hospital beds per 10000 inhabitants and number of doctors per 10000 inhabitants (32 and 27) than countries in Cluster B (22 and 14) and C (12 and 4). Countries in Cluster A recorded a higher level of Chinese imports (US$ 18,000 million) with respect to countries in Cluster B (US$ 7,000 million) and C (US$ 444 million). The three clusters recorded similar percentages of health expenditure over GDP (7% in Cluster A, 6% in Cluster B and 8.6% in Cluster C). The resulting clusters are more related to the closeness to China, rather than to the economic development characteristics of countries.

5.2 Estimation results about the COVID-19 lethality rate

Table 5 presents the Tobit estimation results for all countries in different periods, namely, 60, 180, 360 and 540 days after the first occurrence of the virus in the territory. Tables 6 and 7 present the results for countries of Cluster A and countries of Cluster B[4]. As Tobit and OLS models were applied, Appendix C displays Tobit models for the last periods for comparison purposes. Table 5 presents the general estimation and shows that Cluster A countries recorded a higher lethality rate by 0.007% and 0.006% with respect to Cluster B countries, at 360 and 540 days. The former are those countries that registered COVID-19 cases more rapidly than Cluster B countries. It is worth noting that the difference in the lethality rate between Cluster A and B countries is not significant for the first time periods of 60 and 180 days. While rapidly infected countries registered higher contagion rates than slowly infected countries at the beginning of the pandemic; the lethality rate was only significantly higher later during the pandemic.

It is worth noting that the COVID-19 related variables were time lagged with respect to the lethality rate. The contagion growth rate is lagged by 20 days since once a person became infected, the illness lasted approximately 15 days but longer if conditions worsened (five more days are considered) until death. The stringency index is lagged by 25 days with respect to the lethality rate since once stringency measures were in place, people reduced contact with others. Therefore, we considered that the lethality rate at time t was influenced by stringency measures established at t-25. Our results

[4] For the model of countries in Cluster A, it is worth mentioning that 72 countries out of 78 were considered since six outlier countries were eliminated (Belgium, France, Hungary, Mexico, Peru and Sudan).

Table 6: Estimation of demographic and economic factors affecting the case fatality rate from COVID-19 worldwide

Estimation	TOBIT (marginal effects)			
Variables	60 days	180 days	360 days	540 days
Average lagged contagion growth rate 20 days	0.0535 (0.037)	0.165** (0.08)	-0.192 (0.152)	-0.310 (0.231)
Lagged stringency index 25 days	-9.93e-05 (0.0001)	-2.50e-05 (7.19e-05)	6.32e-05 (6.29e-05)	2.78e-05 (5.94e-05)
Average Government effectiveness index for the last 5 years.	-0.0005*** (0.0002)	-5.61e-05 (8.92e-05)	-0.0002*** (7.20e-05)	-0.0003** (6.72e-05)
Population density for the last available year	-1.32e-06 (3.67e-06)	-4.16e-06** (1.72e-06)	-1.30e-06 (1.52e-06)	-1.10e-06 (1.43e-06)
Percentage of the population over 65 years of age for the last available year	0.0023*** (0.0009)	0.0017*** (0.0004)	0.0007* (0.0003)	0.0008** (0.0003)
Average GDP growth rate for the last 5 years	-0.0025* (0.0015)	-0.0004 (0.0007)	-0.0012* (0.0006)	-0.001* (0.0006)
Average percentage of GDP allocated to health in the last 5 years	0.0019 (0.0016)	0.0019** (0.0008)	0.0004 (0.0006)	0.0003 (0.0006)
Number of hospital beds per 10000 inhabitants for the last available year	-0.0005*** (0.0001)	-0.0003*** (7.40e-05)	-0.0002*** (5.95e-05)	-0.0001** (5.63e-05)
Number of doctors per 10000 inhabitants for the last available year	-0.0001 (0.0002)	-0.0003** (0.0001)	-4.40e-05 (9.54e-05)	-2.43e-05 (8.91e-05)
Logarithm of trade flow (exports) in thousands of dollars between China and its trading partners for the last available year	0.0015 (0.0017)	0.0016** (0.0008)	0.0013* (0.0006)	0.0015** (0.0006)
Group 1 countries	0.0118 (0.0072)	0.002 (0.0037)	0.0073** (0.0034)	0.0058* (0.0031)
Observations	128	120	128	128

Notes: Own elaboration, standard errors in parentheses: ***p<0.01, **p<0.05, *p<0.1

indicate that the contagion growth rate at t-20 significantly increased the lethality rate at time t for the second period (180 day-period) by 0.16%. However, the contagion growth rate was no longer significant for the 360-day and 540-day periods. This also occurred for Clusters A (rapidly infected) and B (slowly infected). The diminished effect of the contagion on the lethality rate reflects the impact of vaccination rollout. During 2020, clinical trials to develop COVID-19 vaccines were in process. By 2021, several vaccines were approved by the Food and Drug Administration (FDA) and they were distributed in developed countries first and then in developing countries. Vaccination marks an important milestone in the worldwide COVID-19 dynamics. An increase of the contagion growth rate did not lead to higher lethality rates a year after the beginning of the pandemic. People were still infected even with the presence of vaccines, but they protected against death. As the World Health Organization (WHO) states, vaccines provide protection against severe illness, hospitalization and death. In addition, the WHO indicates that some evidence shows that vaccines make people less infectious (WHO 2023). Another reason for the lower impact of contagion growth rate on the lethality rate was that the first population segment to get vaccinated were older people, who were recording the higher lethality rates. In general, the stringency index at t-25, although negative for the first periods, did not explain the lethality rate.

An economic variable related to COVID-19 is the trade flow with China. This variable captures the flow of people between China and other countries and therefore the spread of the virus across the globe. Our results show that an increase of the trade flow with China by 1% corresponds to an increase of 0.0014 in the fatality rate from COVID-19 at the global level and in the last three periods for countries in Cluster A. It is not significant for countries in Cluster B since their trade and human mobility relationship is not very strong.

The government effectiveness index measures the perception of the population about government performance and corresponds to a reduction of the lethality rate in both rapidly and slowly infected countries throughout three phases by roughly 0.0004, mean-

Table 7: Estimation of demographic and economic factors affecting the case fatality rate from COVID-19 for Group A countries

Estimation	TOBIT (marginal effects)			
Variables	60 days	180 days	360 days	540 days
Average lagged contagion growth rate 20 days	0.0609 (0.0586)	0.228* (0.121)	-0.225 (0.246)	-0.292 (0.365)
Lagged stringency index 25 days	-0.0003 (0.0002)	3.61e-06 (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)
Average Government effectiveness index for the last 5 years	-0.0007* (0.0003)	-0.0002 (0.0001)	-0.0003** (0.0001)	-0.0004*** (0.0001)
Population density for the last available year	-1.70e-05 (1.81e-05)	0.0028*** (0.0006)	-7.11e-06 (7.56e-06)	-4.99e-06 (7.34e-06)
Percentage of the population over 65 years of age for the last available year	0.0029** (0.0013)	-1.07e-05 (8.68e-06)	0.0009* (0.0005)	0.0011** (0.0005)
Average GDP growth rate for the last 5 years	-0.0016 (0.0024)	-0.0004 (0.0011)	-0.0018** (0.0009)	-0.0012 (0.0009)
Average percentage of GDP allocated to health spending in the last 5 years	0.0033 (0.0027)	0.0019 (0.0012)	-0.0003 (0.001)	-0.0005 (0.001)
Number of hospital beds per 10000 inhabitants for the last available year	-0.0006*** (0.0002)	-0.0005** (0.0001)	-0.0002** (8.33e-05)	-0.0001* (7.95e-05)
Number of doctors per 10000 inhabitants for the last available year	-0.0003 (0.0003)	-0.0002 (0.0002)	-8.35e-05 (0.0001)	-4.49e-05 (0.0001)
Logarithm of trade flow (exports) in thousands of dollars between China and its trading partners for the last available year	0.003 (0.0028)	0.0036*** (0.0013)	0.002* (0.001)	0.0018* (0.0011)
Observations	78	72	78	78

Notes: Own elaboration, standard errors in parentheses: ***p<0.01, **p<0.05, *p<0.1

ing 4 deaths in 10,000 infected people. Liang et al. (2020) obtained that mortality and lethality rate are negatively correlated with government effectiveness. However, the estimated effect of government effectiveness in slowly infected countries was not significant at the beginning of the pandemic. This indicates that the government capacity to establish policies to reduce the effects of the COVID-19 pandemic was not adequate in slowly infected countries. Another explanation could be that effective governments promote an excessive confidence by people, increasing the number of infected people (Toshkov et al. 2021). However, this explanation is discarded as the effect of the average Government effectiveness index for the last 5 year is always associated with a reduction of the lethality rate.

Population density had a limited effect on the worldwide lethality rate, even being non-significant for slowly infected countries. It is negative and significant but still small in magnitude (0.0028) for rapidly infected countries during the second phase. A positively correlated relationship was expected because higher population density is supposed to worsen the spread of the virus, and therefore, the case fatality rate (de Lusignan et al. 2020). Nevertheless, Rodríguez-Pose, Burlina (2021), who also revealed a marginal negative effect of population density, highlight that agglomeration factors are irrelevant to explain excess mortality. They state that excess mortality might be more connected to the interaction and behavior of people, rather than density. Carozzi et al. (2020), state that density might have an immediate influence on outbreaks but have less of an influence on mortality over time.

The main factor that supports the reduction of the COVID-19 lethality rate is the number of hospital beds per 10000 inhabitants. This variable is significant for most models, with exception of Cluster Bat the 360-day period. An increase of 1 bed per 10000 inhabitants reduced the lethality rate, on average, by 0.0003 (3 deaths per 10000 infected people) across periods. This result is in line with the previous literature, indicating that the preexisting sanitary capacity is relevant to control the growth of cases (Rodríguez-Zúñiga et al. 2020, Acosta 2020). Moreover, our results display that the influence of the hospital beds per 10000 inhabitants reduces over time for both rapidly and slowly

Table 8: Estimation of demographic and economic factors affecting the case fatality rate from COVID-19 for Group B countries

Estimation	TOBIT (marginal effects)		OLS	
Variables	60 days	180 days	360 days	540 days
Average lagged contagion growth rate 20 days	0.125*** (0.0413)	0.279* (0.144)	0.0115 (0.275)	0.166 (0.473)
Lagged stringency index 25 days	3.49e-05 (0.0001)	-0.0002 (9.93e-05)	-2.61e-06 (9.10e-05)	-3.02e-05 (8.06e-05)
Average Government effectiveness index for the last 5 years	-0.0002 (0.0002)	-0.0003** (0.0001)	-0.0002 (0.0001)	-0.0003** (0.0001)
Population density for the last available year	-2.12e-08 (2.04e-06)	-1.35e-06 (1.68e-06)	-8.86e-07 (1.54e-06)	-4.71e-07 (1.38e-06)
Percentage of the population over 65 years of age for the last available year	0.0006 (0.0008)	0.0011* (0.0006)	0.0004 (0.0006)	0.0004 (0.0005)
Average GDP growth rate for the last 5 years	-0.0026* (0.0014)	-0.001 (0.001)	-0.0002 (0.001)	-0.0005 (0.0009)
Average percentage of GDP allocated to health spending in the last 5 years	0.0002 (0.0013)	-0.0003 (0.001)	0.0009 (0.001)	0.001 (0.0009)
Number of hospital beds per 10000 inhabitants for the last available year	-0.0005** (0.0002)	-0.0003** (0.0002)	-0.0002 (0.0001)	-0.0002* (0.0001)
Number of doctors per 10000 inhabitants for the last available year	0.0002 (0.0003)	0.0002 (0.0002)	2.96e-05 (0.0002)	0.0001 (0.0002)
Logarithm of trade flow (exports) in thousands of dollars between China and its trading partners for the last available year	-0.0002 (0.0015)	-0.0002 (0.001)	0.0006 (0.0009)	0.0009 (0.0008)
Constants			0.0115 (0.015)	0.0102 (0.0156)
Observations	50	50	50	50
R- squared	—	—	0.193	0.333

Notes: Own elaboration, standard errors in parentheses: ***p<0.01, **p<0.05, *p<0.1

infected countries. The reduction of the lethality rate via hospital beds in the first 60-day period is 0.05% worldwide, 0.06% for rapidly infected countries and 0.05% for slowly infected countries. This reduction decreases over time, to 0.01%, worldwide, 0.01% for rapidly infected countries (Cluster A) and 0.02% for slowly infected countries (Cluster B). These results indicate that the preexisting sanitary capacity was more influential at the beginning of the pandemic.

The percentage of population older than 65 years old is significantly and positively associated to the fatality rate of COVID-19. This result is consistent with King et al. (2020) and Promislow, Anderson (2020), who published that the health risks posed by the virus increase with age. However, its influence on the lethality rate decreases over time from 0.23% at 60 days to 0.08% at 540 days. The percentage of population older than 65 years in Cluster A countries explains their lethality rate in most periods. For Cluster B countries, this variable is significant in only one period. This difference can be explained by the age composition in these countries. On average, the percentage of the population over 65 years of age for Cluster A is 11.35%, while Cluster B, mainly composed of African countries and Caribbean islands, averages 6.79%.

An increase in the average GDP growth over the last five years is associated to a lower lethality rate in most periods of the pandemic. Its effect decreases over time. In the first period, a 1% increase of the GDP growth correlated toa reduction in the lethality rate by 0.25%, whereas the reduction was only 0.1% in the last period. For Cluster A (rapidly infected countries), economic growth corresponds to a reduction of the lethality rate only until the end of the first year of the pandemic. The effect of the economic growth is not significant at the beginning of the pandemic since these rapidly infected countries experienced high rates of infection that potentially contributed to high lethality rates. Other studies conclude a positive relationship between the number of deaths and GDP per capita (Jeanne et al. 2023) because more developed countries are more connected to China and were more affected. In our case, an opposed effect of GDP was obtained since we analyzed different variables; the GDP growth and the lethality rate (number of death cases divided by number of infected people). Rodríguez-Pose, Burlina (2021) conclude that wealthier regions recorded excess mortality because wealthier regions were more accessible by road. The economic growth for countries of Cluster B was significantly associated with lower lethality rates only in the first period of the pandemic. The average percentage of GDP allocated to health spending does not explain the lethality rate in most periods, as obtained also by Khan et al. (2020). The effect of health spending is only significant but positive for the 180-day period. The positive unexpected effect may be explained by the notion that countries with the highest level of contagion and lethality were developed countries with high health spending. Revealed by Chaudhry et al. (2020), it could be explained by more testing and transparency in reporting fatal cases. The correlation rate between the case fatality rate and health spending is 0.1033 for the first period. This correlation coefficient is 0.1331, -0.0137 and 0.0536, for the following periods.

6 Conclusion

By employing Tobit and OLS estimations, this study determined how the underlying conditions of countries in terms of health infrastructure, economic resources and demographic structures influenced theCOVID-19 lethality rate. The results show that both COVID-19 related variables and underlying conditions of countries explain the lethality rate. Risk factors that increase the lethality rate in countries are the contagion growth rate, trade flows with China, the age composition of the population and, to a lesser extent, the population density. Factors that help to reduce the lethality rate are the government effectiveness, the health infrastructure (hospital beds) and, to a lesser extent, the economic growth rate. The contagion growth rate increases the lethality rate, but its influence reduces over time, which may reflect the impact of vaccination. People were still infected once vaccines became more readily available, but the vaccines prevented symptoms from getting worse and causing death. The existing government capacity and structural effectiveness are more important than the conjunctural stringency measures that governments established to reduce the lethality rate. Another structural variable that proved to be an influential factor for the management of the pandemic and reducing the lethality rate is the number of hospital beds per 10000 inhabitants.

It is concluded from the COVID-19 pandemic statistics that while the number of infected people had been increasing across periods, the lethality rate across periods had been decreasing, indicating an improvement in the management of the pandemic by health staff and the role of vaccines. The spread of the virus was not uniform worldwide as some countries registered COVID-19 cases before others. Although the COVID-19 pandemic arrived later to certain countries (slowly infected countries), in most periods, they recorded a higher case fatality rate with respect to rapidly infected countries. While Europe, North America, Oceania and Asia recorded high contagion levels, they recorded lower lethality rates than Latin America and Africa, which recorded low contagion levels and growth.

Using the COVID-19 lethality rate, the contagion growth rate and the number of days that elapsed to register the first confirmed case with respect to China, three clusters of countries are defined: i. Cluster A: Rapidly infected countries with high lethality rate and moderate contagion growth rate, ii. Cluster B: Less rapidly infected countries with very low lethality rate and moderate contagion growth rate, and iii. Cluster C: Slowly infected countries with very high lethality rate and very low contagion growth rate.

Some limitations arose when conducting this research. First, the measurement of confirmed cases and lethal cases records registration problems as some people were asymptomatic (Aliseda 2020) or there was a lack of Polymerase chain reaction (PCR) tests in some territories (Rubino et al. 2020). Regarding the fatal cases, some deaths were registered as severe respiratory infections and not precisely the virus (Parra Saiani et al. 2021). Despite these problems, data from the Blavatnik School of Government in Oxford are available for all countries and correspond to official statistics sent by national governments. For future research, it would be interesting to deeper explore the effect of vaccinations by including the percentage of people with a complete scheme of vaccinations.

Policy implications can be derived from our results. Since the number of doctors per capita had a significant negative correlation with the lethality rate, it is crucial for countries to improve their health system, focusing on health personnel. This action, however, should not be conjunctural, but structural, so the system would be resilient to negative shocks such as the COVID-19 pandemic. In addition, good governance is a key element that facilitates public actions and their effectiveness. Good governance to achieve effectiveness is also a structural feature that must be built constantly, emphasizing on the citizen participation and commonwealth.

References

Abellán A, Aceituno P, Allende A, de Andrés A, Bartumeus F (2020) Una visión global de la pandemia Covid-19: Qué sabemos y qué estamos investigando desde el Csic. *Global Health del CSIC*: 258. https://www.csic.es/sites/default/files/informe_cov19_pti_salud_global_csic_v2_1.pdf

Acosta LD (2020) Capacidad de respuesta frente a la pandemia de COVID-19 en América Latina y el Caribe. *Revista Panamericana de Salud Pública* 44:e109: 8. CrossRef

Aliseda A (2020) Modelos epidemiológicos y COVID-19. Instituto de Investigaciones Filosóficas, Universidad National Autónoma de México, Ciudad de Universidad, México, 1, 7–11

Asfahan S, Shahul A, Chawla G, Dutt N, Niwas R, Gupta N (2020) Early trends of socioeconomic and health indicators influencing case fatality rate of COVID-19 pandemic. *Monaldi Archives for Chest Disease* 9: 451–457. CrossRef

Carozzi F, Provenzano S, Roth S (2020) Urban density and COVID-19. CEP Discussion Paper 1711. London: Centre for Economic Performance

Centers for Disease Control and Prevention (2021) Guía para centros de servicios de día para adultos. https://www.cdc.gov/ncbddd/humandevelopment/covid-19/adult-day-care-service-centers-sp.html

Chaudhry R, Dranitsaris G, Mubashir T, Bartoszko J, Riazi S (2020) A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes. *EClinicalMedicine* 000: 100464. CrossRef

Chisadza C, Clance M, Gupta R (2021) Government effectiveness and the COVID-199 pandemic. *Sustainability* 13. CrossRef

Comisión Económica para América Latina y el Caribe; Naciones Unidas (2020) Los efectos del COVID-19 en el comercio internacional y la logística. *Informe Especial COVID-19*: 24. https://repositorio.cepal.org/handle/11362/45877

de Lusignan S, Dorward J, Correa A, Jones N, Akinyemi O, Amirthalingam G, Andrews N, Byford R, Dabrera G, Elliot A, Ellis J, Ferreira F, Bernal JL, Okusi C, Ramsay M, Sherlock J, Smith G, Williams J, Howsam G, Zambon M, Joy M, Hobbs FD (2020) Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. *The Lancet Infectious Diseases* 3099. CrossRef

Fang W, Wahba S (2020) Urban density is not an enemy in the coronavirus fight: Evidence from China. https://blogs.worldbank.org/sustainablecities/urban-density-not-enemy-coronavirus-fight-evidence-china

GPMB – Global Preparedness Monitoring Board (2019) *Un mundo en peligro: Informe anual sobre preparación mundial para las emergencias sanitarias.* Global Preparedness Monitoring Board. https://www.gpmb.org/docs/librariesprovider17/default-document-library/annual-reports/gpmb-2019-annualreport-en.pdf?sfvrsn=593ede2_3

Guevara-Rosero GC (2022) Clustering countries in the context of the pandemic and underlying conditions. *Studies of Applied Economics* 40. CrossRef

Gujarati DN, Porter DC (2010) *Econometria* (5th ed.). McGraw-Hill, Santa Fe, México. https://fvela.files.wordpress.com/2012/10/econometria-damodar-n-gujarati-5ta-ed.pdf

Han J, Kamber M, Tong H (2022) *Data mining: concepts and techniques* (4th ed.). Morgan Kaufmann, Waltham, MA

Ilardi A, Chieffi S, Iavarone A, Ilardi CR (2021) SARS-CoV-2 in Italy: Population density correlates with morbidity and mortality. *Japanese Journal of Infectious Diseases* 74: 61–64. CrossRef

Javed A (2020) Case fatality rate estimation of COVID-19 for European countries: Turkey's current scenario amidst a global pandemic; comparison of outbreaks with European countries. *Eurasian Journal of Medicine and Oncology* 4: 149–156. CrossRef

Jeanne L, Bourdin S, Nadou F, Noiret G (2023) Economic globalization and the COVID-19 pandemic: Global spread and inequalities. *GeoJournal* 88: 1181–1188

Jinjarak Y, Ahmed R, Nair-Desai S, Xin W, Aizenman J (2020) Accounting for global COVID-19 diffusion patterns, January-April 2020, introduction and overview. *Economics of Disasters and Climate Change* 4: 515–559. CrossRef

Khan JR, Awan N, Islam MM, Muurlink O (2020) Healthcare capacity, health expenditure, and civil society as predictors of COVID-19 case fatalities: A global analysis. *Frontiers in Public Health* 8: 1–10. CrossRef

King JT, Yoon JS, Rentsch CT, Tate JP, Park LS, Kidwai-Khan F, Skanderson M, Hauser RG, Jacobson DA, Erdos J, Cho K, Ramoni R, Gagnon DR, Justice AC (2020) Development and validation of a 30-day mortality index based on pre-existing medical administrative data from 13,323 COVID-19 patients: The Veterans Health Administration COVID-19 (VACO) index. *PLoS ONE* 15: 1–16. CrossRef

Liang LL, Tseng CH, Ho HJ, Wu CY (2020) Covid-19 mortality is negatively associated with test number and government effectiveness. *Scientific Reports* 10: 1–16. CrossRef

Lloyd SP (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28: 129–137. CrossRef

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Le Cam L, Neyman J (eds), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics.* Statistical Laboratory of the University of California, Berkeley, Berkeley, CA

Moreno A, López S, Corcho A (2020) Principales medidas en epidemiología. *Salud Pública de México* 42: 338–348. https://www.scielosp.org/article/ssm/content/raw/?resource_ssm_path=/media/assets/spm/v42n4/2882.pdf

Opertti F, Mesquita-Moreira M (2020) Impacto del coronavirus en el comercio y la integración: ¿qué hacer? – Más Allá de las Fronteras. https://blogs.iadb.org/integracion-comercio/es/coronavirus-comercio-integracion/

PAHO – Pan American Health Organization (2022) Health in the americas in 2022. Retreived from: https://iris.paho.org/bitstream/handle/10665.2/56472/PAHOEIHHA-220024_eng.pdf?sequence=8&isAllowed=y

Park S, Cha Y (2020) The moderating effect of demographic and environmental factors in the spread and mortality rate of COVID-19 during peak and stagnant periods. *Korean Journal of Policy Studies* 35: 77–105. CrossRef

Parra Saiani P, Campo E, Gobo G, Galeotti M (2021) Límites y fallas de los modelos epidemiológicos y predictivos en la epidemia SARS-COV-2 en Italia. In: Federación Española de Sociología (ed), *Impactos Sociales del Covid-19. Miradas desde la Sociología.* Federación Española de Sociología (FES), Madrid, 87–89

Peralta G, Carozzo T, Sierra M, Bu E (2020) Enfermedad por coronavirus (COVID-19): La pandemia según la evidencia actual. *Innovare: Revista de ciencia y tecnología* 9: 15–27. CrossRef

Perone G (2021) The determinants of covid-19 case fatality rate (cfr) in the italian regions and provinces: An analysis of environmental, demographic, and healthcare factors. *Science of the Total Environment* 755: 142523. CrossRef

Promislow DE, Anderson R (2020) A geroscience perspective on COVID-19 mortality. *Journals of Gerontology – Series A Biological Sciences and Medical Sciences* 75: e30–e33. CrossRef

Rocklöv J, Sjödin H (2021) High population densities catalyse the spread of COVID-19. *Journal of Travel Medicine* 27: 1–2. CrossRef

Rodríguez-Zúñiga MJM, Quintana-Aquehua A, Díaz-Lajo VH, Charaja-Coata KS, Becerra-Bonilla WS, Cueva-Tovar K, Valle-Gonzalez GA, Rojas-De-La-Cuba P, Alva-Gutierrez CA, Cerpa-Chacaliaza B, Mendoza-Ticona A (2020) Factores de riesgo asociados a mortalidad en pacientes adultos con neumonía por SARSCoV-2 en un hospital público de Lima, Perú. *Acta Medica Peruana* 37: 437–446. CrossRef

Rodríguez-Pose A, Burlina C (2021) Institutions and the uneven geography of the first wave of the COVID-19 pandemic. *Journal of Regional Science* 61: 728–752. CrossRef

Rubino S, Kelvin N, Bermejo-Martin JF, Kelvin DJ (2020) As COVID-19 cases, deaths and fatality rates surge in Italy, underlying causes require investigation. *Journal of Infection in Developing Countries* 14: 265–267. CrossRef

Sanmartín-Durango D, Henao-Bedoya MA, Valencia-Estupiñán YT, Restrepo-Zea JH (2019) Eficiencia del gasto en salud en la OCDE y ALC: Un análisis envolvente de datos. *Lecturas de Economia*: 41–78. CrossRef

Serikbayeva B, Abdulla K, Oskenbayev Y (2021) State capacity in responding to COVID-19. *International Journal of Public Administration* 44: 920–930. CrossRef

Sorci G, Faivre B, Morand S (2020) Explaining among-country variation in COVID-19 case fatality rate. *Scientific Reports* 10: 1–11. CrossRef

Spyrou MA, Tukhbatova RI, Feldman M, Drath J, Kacki S, Heredia JBD, Arnold S, Sitdikov AG, Castex D, Wahl J, Gazimzyanov IR, Nurgaliev DK, Herbig A, Bos KI, Krause J (2016) Historical Y.pestis genomes reveal the European black death as the source of ancient and modern plague pandemics. *Cell Host and Microbe* 19: 874–881. CrossRef

Toshkov D, Carroll B, Yesilkagit K (2021) Government capacity, societal trust or party preferences: what accounts for the variety of national policy responses to the COVID-19 pandemic in Europe? *Journal of European Public Policy* 29: 1009–1028. CrossRef

Waltenburg MA, Victoroff T, Rose CE, Butterfield M, Jervis RH, Fedak KM, Gabel JA, Feldpausch A, Dunne EM, Austin C, Ahmed FS, Tubach S, Rhea C, Krueger A, Crum DA, Vostok J, Moore MJ, Turabelidze G, Stover D, Donahue M, Edge K, Gutierrez B, Kline KE, Martz N, Rajotte JC, Julian E, Diedhiou A, Radcliffe R, Clayton JL, Ortbahn D, Cummins J, Barbeau B, Murphy J, Darby B, Graff NR, Dostal TKH, Pray

IW, Tillman C, Dittrich MM, Burns-Grant G, Lee S, Spieckerman A, Iqbal K, Griffing SM, Lawson A, Mainzer HM, Bealle AE, Edding E, Arnold KE, Rodriguez T, Merkle S, Pettrone K, Schlanger K, Labar K, Hendricks K, Lasry A, Krishnasamy V, Walke HT, Rose DA, Honein MA (2020) Update: COVID-19 among workers in meat and poultry processing facilities-United States, April-May 2020. *Morbidity and Mortality Weekly Report* 69: 887–892. https://www.cdc.gov/mmwr/volumes/69/wr/mm6927e2.htm

WHO – World Health Organization (2009) Programa mundial de la OMS de investigación de salud pública sobre la gripe. Retrieved from: https://cdn.who.int/media/docs/-default-source/influenza/research-agenda-version-1—languages/research-agenda-esp/-global-influenza-research-agenda-version1_es.pdf?sfvrsn=57903672_1&download=-true#:~:text=programa de investigación-,El Programa mundial de la OMS de investigaciones de salud,brotes de la gripe humana

WHO – World Health Organization (2022) Who health expenditure database. Retrieved in June 2022 from: https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS

WHO – World Health Organization (2023) Coronavirus disease (COVID-19): Vaccines. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-(covid-19)-vaccines

Winskill P, Whittaker C, Walker P, Watson O, Laydon D, Imai N, Cuomo-Dannenburg G, Ainslie K, Baguelin M, Bhatt S, Boonyasiri A, Cattarino L, Ciavarella C, Cooper LV, Coupland H, Cucunuba Z, van Elsland SL, FitzJohn R, Flaxman S, Gaythorpe K, Green W, Hallett T, Hamlet A, Hinsley W, Knock E, John Lees TM, Mishra S, Nedjati-Gilani G, Nouvellet P, Okell L, Parag KV, Thompson HA, Unwin HJT, Vollmer M, Wang Y, Whittles L, Xi X, Ferguson N, Donnelly C, Ghani A (2020) Report 22: Equity in response to the COVID-19 pandemic: An assessment of the direct and indirect impacts on disadvantaged and vulnerable populations in low- and lower middle-income countries. Imperial College COVID-19 response team, Imperial College, London, UK

Woolridge JM (2010) *Introducción a la econometría: Un enfoque moderno* (4th ed.). Cengage Learning, Santa Fe, México

Zevallos JC, Lescano CU (2020) Letalidad y la mortalidad de Covid 19 en 60 países afectados y su impacto en los aspectos demográficos, económicos y de salud. *Revista Medica Herediana* 31: 214–221. CrossRef

Appendix A

Table A.1: Countries Cluster 1

Country	Contagion Rate	Days respect China	CFR	Country	Contagion Rate	Days respect China	CFR
Albania	4,03%	69	1,70%	South Korea	4,24%	22	1,73%
Andorra	7,47%	62	1,02%	Kuwait	6,06%	55	0,57%
Argentina	5,11%	63	2,47%	Lebanon	4,22%	52	1,18%
Armenia	4,46%	61	1,86%	Libya	4,10%	84	1,65%
Austria	4,23%	56	1,89%	Lithuania	4,14%	60	1,64%
Belgium	5,07%	35	2,98%	Latvia	3,86%	62	1,89%
Bulgaria	3,77%	68	4,11%	Morocco	4,36%	62	1,78%
Bahrain	9,11%	55	0,36%	Moldova	4,22%	68	2,13%
Bosnia and Herzegovina	3,58%	65	3,87%	Mexico	5,11%	59	8,83%
Belarus	5,05%	59	0,69%	North Macedonia	4,02%	57	3,08%
Bolivia	3,95%	71	4,65%	Nigeria	4,03%	59	1,22%
Brazil	5,77%	57	2,43%	Nicaragua	5,11%	79	2,68%
Brunei	4,13%	69	1,60%	Netherlands	5,48%	58	1,43%
Canada	4,00%	23	2,68%	Norway	4,68%	57	0,89%
Switzerland	5,87%	56	1,81%	Nepal	4,09%	25	0,73%
Chile	4,72%	54	2,51%	Pakistan	4,43%	56	2,20%
China	7,41%	0	5,33%	Panama	5,56%	70	1,72%
Cote d'Ivoire	4,22%	71	0,57%	Peru	5,51%	66	9,28%
Colombia	5,13%	66	2,65%	Philippines	4,42%	30	1,99%
Costa Rica	4,54%	66	1,37%	Poland	5,31%	64	2,57%
Czechia	4,23%	61	1,66%	Portugal	4,43%	62	2,02%
Germany	5,26%	27	2,41%	Paraguay	4,34%	68	1,99%
Denmark	4,65%	58	1,13%	Qatar	6,30%	60	0,16%
Dominican Republic	4,41%	61	1,29%	Romania	4,67%	57	2,55%
Algeria	4,06%	56	2,65%	Russia	4,88%	31	1,86%
Ecuador	3,69%	61	5,61%	Saudi Arabia	4,99%	62	1,72%
Egypt	4,78%	45	5,70%	Sudan	3,49%	74	6,24%
Spain	5,37%	32	2,16%	El Salvador	3,75%	79	3,13%
Estonia	4,40%	58	0,93%	San Marino	4,23%	60	2,06%
Ethiopia	4,37%	73	1,46%	Serbia	4,94%	66	0,96%
France	4,86%	24	2,39%	Slovakia	4,47%	66	1,23%
United Kingdom	4,97%	31	2,68%	Slovenia	4,13%	65	2,02%
Georgia	4,38%	57	1,28%	Sweden	4,72%	32	2,02%
Greece	4,47%	57	3,51%	Eswatini	3,51%	74	3,83%
Guatemala	4,07%	74	3,62%	Syria	3,72%	82	6,68%
Croatia	4,18%	56	2,26%	Togo	4,36%	66	1,23%
Hungary	3,94%	64	3,51%	Tunisia	4,20%	64	3,43%
Indonesia	4,57%	62	2,70%	Turkey	6,14%	71	1,05%
India	5,98%	30	1,44%	Uganda	5,03%	81	0,82%
Ireland	4,57%	60	1,93%	Ukraine	4,95%	63	1,97%
Iran	4,96%	50	3,90%	United States	5,79%	22	1,68%
Iraq	4,89%	55	2,01%	Uzbekistan	4,51%	75	0,77%
Israel	4,58%	52	0,74%	Venezuela	4,22%	74	0,98%
Italy	5,94%	31	3,47%	Yemen	3,48%	101	19,57%
Jordan	5,52%	63	1,21%	South Africa	4,97%	65	3,30%
Japan	3,74%	22	1,38%	Zimbabwe	3,67%	80	4,12%
Kenya	4,05%	73	1,72%				

Table A.2: Countries Cluster 2

Country	Contagion Rate	Days respect China	CFR	Country	Contagion Rate	Days respect China	CFR
Aruba	2,74%	73	0,95%	Laos	1,10%	84	0,00%
Afghanistan	3,03%	55	4,37%	Liberia	2,12%	77	4,20%
Angola	3,18%	80	2,43%	Saint Lucia	2,78%	74	1,18%
Anguilla	0,75%	88	0,00%	Liechtenstein	2,74%	64	2,08%
United Arab Emirates	3,50%	29	0,29%	Sri Lanka	3,75%	27	0,49%
Antigua and Barbuda	2,45%	73	2,48%	Lesotho	3,47%	134	2,96%
Australia	2,90%	26	3,16%	Luxembourg	3,91%	60	1,16%
Azerbaijan	3,63%	61	1,37%	Macao	1,44%	22	0,00%
Burundi	2,48%	91	0,23%	Monaco	2,80%	60	1,21%
Benin	3,17%	76	1,25%	Madagascar	3,20%	80	1,55%
Bonaire Sint Eustatius and Saba	1,52%	93	1,53%	Maldives	2,69%	68	0,31%
Burkina Faso	3,73%	70	1,18%	Mali	2,89%	85	3,96%
Bangladesh	3,90%	68	1,54%	Malta	2,85%	67	1,39%
Bahamas	3,12%	76	2,14%	Myanmar	3,06%	87	2,25%
Belize	3,09%	83	2,55%	Montenegro	3,93%	77	1,35%
Bermuda	2,18%	79	1,63%	Mongolia	2,99%	70	0,06%
Barbados	2,53%	77	1,09%	Mozambique	3,70%	82	1,12%
Bhutan	2,23%	66	0,12%	Mauritania	3,29%	74	2,55%
Botswana	3,13%	90	1,32%	Montserrat	1,63%	78	5,00%
Central African Republic	3,39%	75	1,25%	Mauritius	1,92%	78	1,45%
Cameroon	3,68%	66	1,54%	Malawi	3,02%	93	3,33%
Democratic Republic of Congo	3,59%	71	2,69%	Malaysia	3,29%	25	0,37%
Congo	3,86%	75	1,40%	Namibia	3,13%	74	1,10%
Comoros	3,37%	121	3,81%	New Caledonia	1,20%	79	0,00%
Cape Verde	3,90%	80	0,97%	Niger	3,15%	80	3,74%
Cuba	3,04%	72	0,62%	New Zealand	2,88%	59	1,10%
Curacao	2,88%	74	0,46%	Oman	3,64%	55	1,12%
Cayman Islands	2,21%	73	0,44%	Kosovo	3,58%	74	2,24%
Cyprus	3,41%	70	0,64%	Papua New Guinea	2,68%	80	1,15%
Djibouti	3,45%	78	1,01%	Palestine	3,44%	65	1,11%
Dominica	1,99%	82	0,00%	French Polynesia	2,96%	73	0,76%
Eritrea	3,07%	81	0,23%	Rwanda	3,83%	74	1,37%
Finland	3,54%	29	1,61%	Senegal	3,61%	62	2,53%
Fiji	1,42%	79	3,03%	Singapore	3,70%	23	0,05%
Falkland Islands	1,60%	95	0,00%	Sierra Leone	2,79%	91	1,99%
Faeroe Islands	2,57%	64	0,15%	Somalia	3,07%	76	3,83%
Gabon	3,37%	74	0,57%	Saint Pierre and Miquelon	1,20%	96	0,00%
Ghana	3,44%	74	0,76%	South Sudan	3,88%	96	1,07%
Gibraltar	3,09%	64	2,19%	Sao Tome and Principe	3,39%	97	1,52%
Guinea	3,38%	73	0,57%	Suriname	3,38%	74	1,95%
Gambia	2,89%	77	3,10%	Seychelles	2,23%	75	0,47%
Guinea-Bissau	3,09%	85	1,55%	Turks and Caicos Islands	2,04%	88	0,74%
Equatorial Guinea	3,54%	75	1,51%	Chad	2,97%	79	3,57%
Grenada	2,75%	82	0,65%	Thailand	2,40%	22	0,60%

Country	Contagion Rate	Days respect China	CFR	Country	Contagion Rate	Days respect China	CFR
Greenland	1,26%	76	0,00%	Tajikistan	2,45%	122	0,66%
Guyana	3,75%	72	2,28%	Timor	1,76%	82	0,00%
Hong Kong	2,77%	23	1,70%	Trinidad and Tobago	3,42%	74	1,81%
Honduras	3,71%	71	2,45%	Taiwan	2,37%	22	0,82%
Haiti	3,12%	80	1,98%	Tanzania	2,49%	76	4,13%
Isle of Man	2,83%	80	1,99%	Uruguay	3,35%	73	1,02%
Iceland	3,24%	59	0,48%	Vatican	1,54%	66	0,00%
Jamaica	3,61%	71	1,76%	Saint Vincent and the Grenadines	2,55%	74	0,48%
Kazakhstan	3,84%	73	1,29%	British Virgin Islands	1,49%	88	0,65%
Kyrgyzstan	3,63%	78	1,70%	Vietnam	2,33%	23	2,28%
Cambodia	2,63%	27	0,00%	Zambia	3,82%	78	1,36%
Saint Kitts and Nevis	1,28%	85	0,00%				

Table A.3: Countries Cluster 3

Country	Contagion Rate	Days respect China	CFR	Country	Contagion Rate	Days respect China	CFR
Micronesia (country)	0,00%	387	0,00%	Vanuatu	0,69%	315	16,67%
Marshall Islands	0,74%	302	0,00%	Wallis and Futuna	2,55%	293	1,54%
Saint Helena	0,25%	251	0,00%	Samoa	0,44%	323	0,00%
Solomon Islands	0,86%	286	0,00%				

Appendix B

Table B.1: Descriptive statistics by clusters

	Cluster A				Cluster B				Cluster C			
Variable	Min.	Max.	Std. Dev	Mean	Min.	Max.	Std. Dev	Mean	Min.	Max.	Std. Dev	Mean
Case fatality rate	0.0	0.2	0.02	0.02	0	0.1	0.01	0.02	0.0	0.0	0.0	0.0
Lagged contagion growth rate at 540 days (t-20)	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Stringency index at 540 days (t-25)	2.8	86.1	16.8	51.4	11.1	93.5	18.4	51.5	0.0	0.0	0.0	0.0
Government effectiveness	0.0	95.0	22.8	49.7	7.5	99.6	24.1	44.7	0.0	0.0	0.0	0.0
Percentage of population older than 65 years old	1.5	28.0	6.8	11.4	1.2	22.1	5.0	6.8	3.6	4.9	0.5	4.1
Population density	3.8	2017.3	236.0	154.4	0.1	20777.5	2951.0	757.0	23.3	324.5	113.7	120.4
GDP growth (average of the 5 last years)	-10.4	10.1	2.5	2.9	-10.8	7.4	2.9	3.0	1.9	3.6	0.6	3.0
Current health expenditure (average of the 5 last years)	2.4	16.9	2.4	7.1	1.7	17.4	2.5	5.8	3.3	17.2	5.3	8.7
Hospital beds per 10000 inhabitants	3.0	129.8	26.2	32.9	1.0	80.0	16.5	21.8	10.0	14.0	2.0	12.0
Number of doctors per 10000 inhabitants	0.8	79.3	18.2	26.9	0.2	83.0	17.3	14.3	1.8	5.8	1.8	3.6
Chinese exports to each country i in billions of dollars for the last available year	2.4	418584.2	48763.0	18422.8	0.006	279616.8	30447.2	7298.5	0.7	2384.7	868.4	444.6

Appendix C

Table C.1: TOBIT estimation of demographic and economic factors affecting the case fatality rate from COVID-19 for the countries of Group 2 for 360 and 540 days

	TOBIT (marginal effects)	
Variables	360 days	540 days
Average lagged contagion growth rate 20 days	0.0628 (0.231)	0.153 (0.384)
Lagged stringency index 25 days	-3.00e-06 (7.53e-05)	-2.78e-05 (6.54e-05)
Average Government effectiveness index for the last 5 years	-0.0002* (9.38e-05)	-0.0003*** (8.21e-05)
Percentage of the population over 65 years of age for the last available year	0.0004 (0.0005)	0.0003 (0.0004)
Population density for the last available year	-7.92e-07 (1.27e-06)	-4.34e-07 (1.12e-06)
Average GDP growth rate for the last 5 years	-0.0003 (0.0008)	-0.0005 (0.0007)
Average percentage of GDP allocated to health spending in the last 5 years	0.0008 (0.0008)	0.001 (0.0007)
Number of hospital beds per 10000 inhabitants for the last available year	-0.0002 (0.0001)	-0.0002** (9.77e-05)
Number of doctors per 10000 inhabitants for the last available year	1.59e-05 (0.0002)	0.0001 (0.0002)
Logarithm of trade flow (exports) in thousands of dollars between China and its trading partners for the last available year	0.0005 (0.0007)	0.0008 (0.0006)
Observations	50	50

REGION
The Journal of ERSA
Powered by WU

Volume 11, Number 1, 2024, 55–78
DOI: 10.18335/region.v11i1.523

journal homepage: region.ersa.org
ISSN: 2409-5370

Bioeconomy firms and where to find them

Lukas Kriesch[1], **Sebastian Losacker**[1]

[1] Justus-Liebig-University Giessen, Giessen, Germany

Received: 15 November 2023/Accepted: 7 March 2024

Abstract. The bioeconomy represents a transformative approach to economic development and sustainability by harnessing biological resources and knowledge to produce goods, services, and energy while reducing dependence on non-renewable resources. In order to understand and support the bioeconomy, scholars and policymakers rely on an accurate measurement and monitoring of bio-based economic activities. However, existing statistical frameworks and industry classifications often fall short in capturing the unique characteristics of the bioeconomy. This article addresses this challenge by developing a methodological approach for comprehensive measurement and mapping of bio-based economic activities. We build a novel data set of bioeconomy firms in Germany using web-mining and machine learning techniques. This data set enables detailed analysis of bio-based economic activities, providing valuable insights into the spatial organization of the bioeconomy. The paper demonstrates the applicability of the data set by testing several hypotheses about the bioeconomy. Our research contributes to a better understanding of the bioeconomy's regional impacts and offers a valuable resource for policymakers and researchers interested in understanding the geography of bio-based economic activities. We make an aggregated version of the data set freely available online.

Key words: bioeconomy, web mining, natural language processing, regional data

1 Introduction

The bioeconomy represents a paradigm shift in our approach to economic development and sustainability. It encompasses the sustainable utilization of biological resources, such as plants, microorganisms, and biomass to produce a wide range of goods, services, and energy. The bioeconomy not only recognizes the potential of biological resources to meet our growing needs, but also acknowledges the need to reduce our dependence on non-renewable resources and to mitigate environmental impacts. By integrating novel sustainable technologies, innovative processes, and principles of circularity, the bioeconomy could offer a pathway towards a more sustainable and resilient future (Aguilar et al. 2018, Befort 2023, Bugge et al. 2016, Patermann, Aguilar 2021). Based on this vision, many countries have implemented a range of bioeconomy policies and strategies aiming to foster sustainable development (Prochaska, Schiller 2021, Proestou et al. 2023, Vogelpohl, Töller 2021).

The bioeconomy is also a promising concept for regional economies, as it offers regions the opportunity to diversify their economic base, foster innovation, and create new employment opportunities. By capitalizing on local biological resources and knowledge capabilities, regions can develop specialized clusters and value chains that leverage their unique ecological assets and knowledge bases (Kamath et al. 2023, Laasonen 2023, Martin et al. 2023, Morales, Dahlström 2022). In this context, the bioeconomy presents a viable pathway for regions to transition toward more sustainable economies. It can promote the adoption of sustainable practices, such as resource efficiency, waste valorization, and the circular economy, leading to reduced ecological footprints and enhanced regional sustainability.

However, it is important to note that the positive impacts of the bioeconomy, as well as its potential to foster regional development and sustainability, often remain speculative and are not guaranteed. The narrative of the bioeconomy as a universally beneficial approach is predominantly advocated by policymakers and related stakeholders who are keen on promoting its adoption. Yet, there is a growing body of scholarly work that raises critical questions regarding the assumed benefits of the bioeconomy (Allain et al. 2022, Bauer 2018, Bringezu et al. 2021, Friedrich et al. 2021). In that regard, economic geographers and regional scientists play a crucial role in analyzing the spatial dynamics of the bioeconomy, assessing its impacts on regional economies and informing policy interventions that foster sustainable regional development. Hence, scholars in economic geography and regional science can contribute to a better understanding of regional structural change and regional sustainability transitions towards a future bio-based economy.

Against this background, accurate tracking of bioeconomy activities is essential, not only for research purposes, but also for policymakers seeking to design effective strategies and place-based policies. Understanding the size, scope, and trends of bio-based economic activities provides policymakers with crucial insights into the bioeconomy's contribution to regional and national economies, job creation, and environmental sustainability. It enables them to identify emerging sectors, target support measures, and assess the effectiveness of policy interventions (El-Chichakli et al. 2016, Wesseler, von Braun 2017). However, measuring bio-based economic activities presents significant challenges. Traditional economic indicators often fail to capture the unique characteristics of the bioeconomy, such as the integration of biological resources and the circularity and sustainability of economic processes. Moreover, existing statistical frameworks, industry classifications and databases may lack comprehensive data on bioeconomy-related activities, making it difficult to obtain a complete and accurate picture. The multidimensional nature of the bioeconomy, spanning various sectors and encompassing both tangible and intangible elements, further complicates measurement efforts (Fischer et al. 2024, Losacker et al. 2023b, Ronzon et al. 2017, Wydra 2020).

In this paper, we contribute to solving these issues. The aim of this paper is to develop a methodological approach that allows a comprehensive measurement of bio-based economic activities. In that vein, we also aim to unveil the geography of bio-based economic activities. To this end, we build a unique dataset that enables us to identify and map bioeconomy firms in Germany. The dataset is based on a novel web-mining approach developed by Kriesch (2023). This dataset uses the open-source web repository CommonCrawl to identify German company websites and has proven to be a valuable database for spatial research. From this data, we identify bioeconomy firms using a combination of different natural language processing techniques, utilizing the semantic capabilities of modern transformer models (Reimers, Gurevych 2019, Vaswani et al. 2017). Our empirical approach allows for a detailed analysis of the economic activities of bio-based firms. That is to say, we are able to assess firms' technological capabilities and we can understand in which domains firms operate. In short, we establish a novel data source for monitoring the bioeconomy, overcoming several issues researchers and practitioners usually face when trying to measure bio-based economic activities. We test several hypotheses on the bioeconomy to validate our dataset and to demonstrate its applicability for future research, which is commonly done when introducing new methods or data to regional research (Abbasiharofteh et al. 2023, Ozgun, Broekel 2022). We make an aggregated version of our dataset freely accessible for fellow researchers, enabling further analyses and contributions to regional bioeconomy studies.

The remainder of this paper is organized as follows. In Section 2, we will introduce ideas behind the bioeconomy concept and review empirical findings that help to understand the geography of bio-based economic activities. This allows us to derive a couple of hypotheses about the bioeconomy and its geography. In Section 3, we explain our methodological approach in detail. We present our main results in Section 4. Section 5 concludes.

2 Literature review: What do we know about the bioeconomy?

2.1 The idea of a bio-based economy

The idea of a bioeconomy has emerged as a response to the need for sustainable development and the challenges posed by resource scarcity, climate change and environmental degradation. It refers to the (sustainable) utilization of biological resources to produce goods, services, and energy, and it encompasses diverse sectors such as agriculture, forestry, fisheries, biotechnology, and renewable energy. The use of bio-based products and technologies, however, is sought to span all economic sectors. The purpose of the bioeconomy is to shift from a linear economic model to a circular and regenerative approach, where biological resources are efficiently and responsibly managed (Aguilar et al. 2018, Allain et al. 2022, Befort 2023, Bugge et al. 2016, Patermann, Aguilar 2021).

For a transition towards a bio-based economy, the role of innovation emerges as paramount. Innovation, defined as any novel economic activity, is crucial to transition from a fossil economy to a bioeconomy (Befort 2023). As such, the bioeconomy encompasses a spectrum of innovation types, from drop-in solutions and bio-based substitutes to more transformative bio-based innovations that reshape socio-technical systems and redefine production networks (Befort 2023, Giurca, Befort 2023, Kuckertz et al. 2020, Losacker et al. 2023b). Nevertheless, the extent to which the bioeconomy and its innovations genuinely contribute to a more sustainable future remains somewhat ambiguous (Allain et al. 2022, Bauer 2018, Bringezu et al. 2021, Friedrich et al. 2021).

In the evolving discourse on the bioeconomy, there exists a multifaceted understanding of its implications and potential trajectories. In this regard, Bugge et al. (2016) delineate three distinct visions that encapsulate the breadth of scholarly perspectives on the bioeconomy. Firstly, the *biotechnology vision* underscores the pivotal role of biotechnology research, emphasizing its application and commercialization across diverse economic sectors. This vision is rooted in the belief that technological advancements and scientific progress can drive economic growth and innovation. Secondly, the *bio-resource vision* is anchored in the processing and enhancement of biological raw materials. It envisions a future where new value chains are established, leveraging the inherent potential of biological resources. Lastly, the *bio-ecology vision* emerges as a sustainability-centric perspective. It accentuates the importance of ecological processes that optimize energy and nutrient utilization, champion biodiversity, and caution against the pitfalls of monocultures and soil degradation. While expectations about actual sustainability outcomes of a future bioeconomy may vary according to these visions, there exists a broad consensus in the scholarly debate that the bioeconomy encompasses a wide array of industries and sectors (see above). This includes both traditional, low-tech goods and services in sectors such as forestry and agriculture, as well as more complex, knowledge-intensive economic activities like R&D in biotechnologies. In other words, the economic activities that can be associated with the bioeconomy are very diverse and thus difficult to track – and so is their geography.

2.2 The geography of bio-based economic activities

The bioeconomy emerges as a potent avenue for regional development, potentially enabling regional diversification, fostering regional innovation, and generating local employment opportunities. By utilizing local biological resources and leveraging regional knowledge capabilities, regions can establish specialized clusters and value chains that capitalize on their unique ecological assets and knowledge bases. In regional research, the bioeconomy has therefore received pronounced attention in recent years. For example, several researchers have delved into regional structural change and sustainability transitions toward a bio-based regional economy (Halonen et al. 2022, Laasonen 2023, Sanz-Hernández et al. 2019). Related studies have investigated innovative bio-clusters and regional innovation systems centered on biotechnologies (Abbasiharofteh, Broekel 2020, Heimeriks, Boschma 2014, Kamath et al. 2023), innovation networks (Bauer et al. 2018), and regional bioeconomy policies and strategies (Andersson, Grundel 2021). Additionally, there is increasing research interest in regional path creation and the pivotal role of actors and agency in propelling regional bioeconomy transitions (Martin et al. 2023, Morales, Dahlström 2022, Steinböck, Trippl 2023). Drawing from previous research findings on the bioeconomy, and complemented by insights from regional economics and economic geography, we can deduce several hypotheses concerning the geography of bio-based economic activities. In the following, we will enumerate these hypotheses and elucidate the rationale behind their formulation.

The bioeconomy, with its strong reliance on biomass and bio-based resources, raises the question of where bioeconomy firms tend to concentrate. There are compelling reasons to believe that bioeconomy enterprises tend to gravitate towards rural regions, resulting in the formation of rural bioeconomy clusters. These clusters may emerge as 'agricultural agglomerations' or 'Marshallian bio-districts' (Hermans 2021). Firstly, rural areas offer abundant biomass resources, including forests and agricultural products, providing a competitive advantage to bioeconomy firms reliant on biomass feedstocks. Secondly, rural regions often have established synergies with traditional industries like agriculture and forestry, offering infrastructure, knowledge, and expertise that bioeconomy firms can leverage for collaboration and innovation. Lastly, policy and regional development initiatives play a significant role in attracting bioeconomy firms to rural areas through financial incentives, grants, and supportive frameworks. These policies are often aimed at promoting rural development and can create a favorable investment climate for bioeconomy activities (Prochaska, Schiller 2024, Haarich, Kirchmayr-Novak Haarich, Kirchmayr-Novak). Based on these arguments, we assume that bioeconomy firms generally concentrate in rural areas, a hypothesis that is also supported by related empirical studies (Lasarte Lopez et al. 2023, Refsgaard et al. 2021).

Hypothesis 1 : *Given that the bioeconomy strongly relies on biomass and bio-based resources, bioeconomy firms concentrate in rural areas.*

In contrast to this first proposition, we argue that complex bioeconomy activities, such as those in biotechnology, concentrate in urban regions. Several arguments from the geography of innovation literature support this claim, building on core reasonings about agglomeration economies (Asheim et al. 2016, Broekel et al. 2023, Losacker et al. 2023a). Firstly, urban areas often provide a conducive environment for innovation and knowledge exchange. The concentration of universities, research institutions, and diverse talent pools in urban regions fosters collaboration, networking, and exchange of ideas. The availability of human capital and a strong regional innovation system attracts and supports the development of innovative bioeconomy activities. Secondly, urban regions typically have better access to specialized infrastructure and resources that are essential for cutting-edge research and development. Research facilities, laboratories, and technology parks are more prevalent in urban areas, providing infrastructure and equipment necessary for complex bioeconomy activities. Moreover, urban areas offer advanced transportation networks, communication systems, and logistical support, facilitating the movement of goods, services, and knowledge-intensive activities required for many innovative bioeconomy firms. Thirdly, urban regions often provide larger market opportunities and a diverse customer base. The concentration of various industries, markets, and consumers in urban areas creates a significant potential market for innovative bioeconomy products and services (Cooke 2002, Hermans 2018). Existing research supports this view, indicating that urban areas often host a higher concentration of complex bioeconomy activities (Ehrenfeld, Kropfhäußer 2017). In summary, while the hypothesis on bioeconomy firms concentrating in rural areas due to the strong reliance on biomass remains valid, there are additional reasons to claim that complex innovative bioeconomy activities, e.g., in biotechnology, concentrate in urban regions.

Hypothesis 2 : *Bioeconomy innovations and high-tech activities concentrate in urban areas.*

Next, we argue that economic activities centered on bio-based processes and biomass, such as activities in forestry or agriculture, typically locate in close proximity to their primary biomass feedstocks, a locational feature typical for bioeconomy clusters (Hermans 2021). This is underpinned by several reasons aligned with the basic principles of Weberian location theory. Firstly, being near the source of raw materials minimizes transportation costs, ensuring that the feedstock remains cost-effective for production or processing. Secondly, proximity to biomass sources ensures a consistent and timely supply, reducing potential downtimes or disruptions in the production process. Furthermore, being close to the source often means fresher inputs, which can be crucial for certain bio-based processes that rely on the quality and freshness of biomass. Lastly, such co-location fosters synergies with local agricultural or forestry sectors, promoting integrated value chains and facilitating efficient resource utilization. This geographical alignment between bio-based activities and their feedstock sources is not only economically prudent but also aligns with principles of sustainable production and consumption. The co-location of bioeconomy firms to biomass feedstocks is evident from several regional case studies on the bioeconomy (Martin et al. 2023, Martin, Coenen 2014, Ramirez 2021).

Hypothesis 3 : *Economic activities centered on bio-based processes and biomass locate in close proximity to their primary biomass feedstocks.*

We acknowledge that these hypotheses are somewhat generic. Nevertheless, their primary function is to highlight and showcase the novel dataset we have constructed in this paper, constituting the core contribution of our work.

3 Methods: Using web text data to map the bioeconomy

3.1 Issues in measuring the bioeconomy

We argue that previous attempts of measuring the bioeconomy are insufficient, mainly because the bioeconomy spans multiple sectors and can therefore not be captured using traditional methods or indicators. One of the challenges in measuring bio-based economic activities lies in the inadequacy of traditional statistical classifications, such as NACE and SIC codes or other sectoral classifications of economic activities, to fully capture the diverse nature of the bioeconomy. These classifications were primarily designed to categorize economic activities based on conventional industry sectors, often overlooking the unique characteristics and interconnections of bio-based economic activities. The bioeconomy, by its very nature, cuts across multiple sectors and involves a wide range of activities that may not neatly fit into traditional sectoral boundaries. For example, the bioeconomy includes sectors like biotechnology that span across different industries, combining elements of agriculture, manufacturing, and healthcare. It also encompasses activities like bio-energy production, bio-refineries, and bio-materials development, which do not align with conventional industry classifications (Jander, Grundmann 2019, Ronzon et al. 2017, Wesseler, von Braun 2017). Furthermore, the bioeconomy is characterized by innovation, continuous technological advancements, and the emergence of new value chains. Statistical classifications tend to be static and may struggle to keep pace with the dynamic and rapidly evolving nature of the bioeconomy. This dynamic nature often results in novel business models, cross-sector collaborations, and disruptive innovations that may not be adequately captured by existing statistical frameworks. These limitations are not only valid for measuring economic activities, but also for measuring innovation activities and knowledge generation where traditional indicators (e.g., patent data) also rely on statistical classifications that are not able to fully capture all parts of the bioeconomy (Fischer et al. 2024, Losacker et al. 2023b, Wydra 2020).

Prior efforts to gauge the bioeconomy have often relied on 'sector shares,' wherein researchers devise methodologies to determine the proportion of bio-based activities within traditional sector classifications like NACE (Lasarte Lopez et al. 2023, Ronzon et al. 2017). While this approach may be reasonable when assessing the bioeconomy at a national level, it is susceptible to what statisticians term an ecological fallacy when examined from a geographical standpoint. This fallacy involves making inferences about the characteristics of individual units (such as firms or regions) based on generalizations about the entire group (such as a nation). In many instances, researchers assume that, for example, 20% of the economic activity (value added, employment, etc.) in a specific sector is part of the bioeconomy. While this assumption may hold true on an aggregate level, it can lead to erroneous conclusions when applied to the firm or regional level. The share of bio-based activities within the sector may vary significantly across regions, being either notably higher or lower.

To overcome these limitations, researchers and policymakers need to explore alternative approaches that go beyond traditional statistical classifications. These approaches include adopting more flexible and adaptive frameworks that can capture the multidimensional and cross-sectoral aspects of the bioeconomy. Such frameworks may involve the development of new classification systems, the use of hybrid models that combine qualitative and quantitative data, and the integration of emerging indicators that reflect the unique characteristics of bio-based economic activities more adequately. In the next two sections (3.2 and 3.3), we propose an alternative way of measuring the bioeconomy. That is to say, we use a web-mining approach to retrieve information on firms from their website texts. Based on this text data, we employ machine learning techniques to identify, classify, and map bio-based economic activities.

3.2 A web-mining approach to identify bio-based economic activities in Germany

As data source we use website texts from German companies identified by Kriesch (2023). Our research focuses on the identification of bio-based economic activities in Germany, as Germany has emerged as a leading proponent of bioeconomy policies, recognizing its potential to drive sustainable economic growth and address environmental challenges (Imbert et al. 2017, Prochaska, Schiller 2021). At the regional level, Germany has implemented a range of policies and initiatives to promote the bioeconomy, supporting regional collaborations between businesses, research institutions, and policymakers to advance bio-based innovations. These policies aim to create favorable conditions for regional businesses to invest in bioeconomy activities, develop sustainable value chains, and contribute to regional economic development.

Web mining has witnessed significant advancements in recent years, primarily driven by the rise of natural language processing (NLP) techniques and the increasing digitization of data. Accordingly, the integration of web text data has become a vital complement to traditional data sources. In particular, unconventional information that may not be captured by traditional sources can be unveiled by means of web mining. Among other topics, companies use their websites to display information about products and services, their orientation and beliefs, strategies and relations with other companies (Gök et al. 2015). The dataset used in this paper, developed by Kriesch (2023), consists of 678,381 companies and their corresponding website texts. We updated the original dataset again in July 2023 to ensure that we have up-to-date information. We utilized geocoding techniques to map websites to geographical locations based on the information provided in their imprints. The address information was extracted using a fine-tuned named entity recognition model. We then used address geocoding to convert the address information into coordinates. In cases where a single company operates multiple domains, we attributed the content from each domain to the geographical location specified in its imprint. We extracted the HTML code of the first 25 subpages located on the landing page for each company domain, disregarding URLs predominantly comprised of machine-generated content and general legal information (e.g., imprint, cookie-policy, terms and conditions). Following Kinne, Lenz (2021), we argue that text located closer to the front or beginning of a domain is more likely to present general information about the whole company, while text positioned further back or towards the end tends to contain more specific details. Following the scraping process, the dataset comprises 9,601,260 subpages.

Text pre-processing plays a crucial role in converting raw HTML code into meaningful text data. Particularly, web data often contains a substantial amount of low quality and machine-generated content that may not be suitable for accurately predicting a company's capabilities. Hence, it is essential to employ appropriate data filtering and

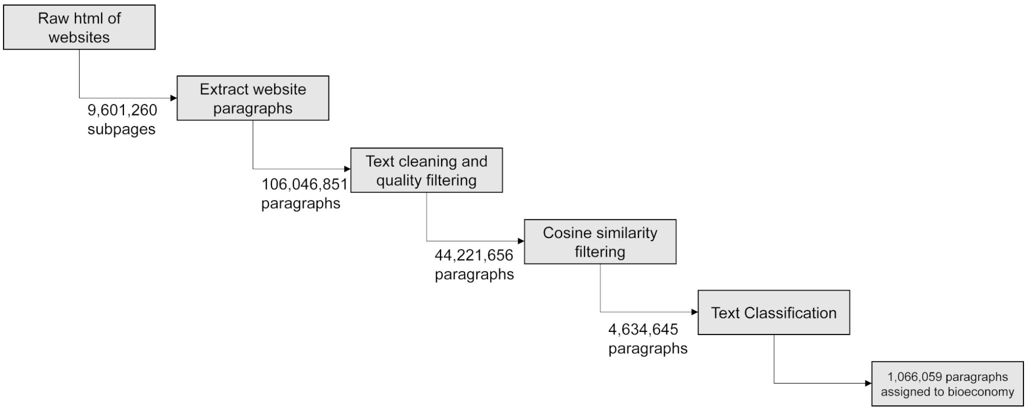

Raw html of websites
9,601,260 subpages
Extract website paragraphs
106,046,851 paragraphs
Text cleaning and quality filtering
44,221,656 paragraphs
Cosine similarity filtering
4,634,645 paragraphs
Text Classification
1,066,059 paragraphs assigned to bioeconomy

Figure 1: Flowchart of website text data processing

pre-processing techniques to mitigate the impact of irrelevant or misleading content on the predictive performance of the model. We illustrate the pre-processing procedure in Figure 1. In steps one and two, we extracted the main text from the HTML code, segmented the text into paragraphs and removed unwanted content like menus, headers, footers or advertisements. In the third filtering step, we applied further quality filters at the paragraph level to extract coherent and contextually well-embedded text. For this purpose, we use slightly adapted quality filtering heuristics developed by Rae et al. (2021). These heuristics have proven to be effective for the training of large language models and can be applied to prepare our dataset for analysis. Detailed information on the modifications made to these heuristics can be found in Appendix A.

We conducted a semantic search to identify paragraphs related to the bioeconomy. Therefore, we use a Sentence-BERT (SBERT) model to transform each of the extracted paragraphs into a complex numerical vector. SBERT models have demonstrated remarkable efficiency in leveraging the semantic knowledge of pre-trained transformer models, particularly when applied to downstream tasks such as semantic search or clustering. Their ability to capture and utilize semantic meanings has led to notable advancements in these specific applications, improving both accuracy and efficiency (Reimers, Gurevych 2019). We also embedded different keywords referring to the bioeconomy, as detailed in Appendix B, using the same SBERT model (German_Semantic_STS_V2). The choice of keywords for the semantic search and the methodology for manual annotation were thoroughly deliberated during a workshop involving bioeconomy experts. This collaborative session proved instrumental in gaining a nuanced understanding of what truly pertains to the bioeconomy. Unlike keyword-based searches that rely solely on matching specific terms, a semantic search employs advanced language understanding techniques to identify related and semantically similar concepts. This expanded scope of a semantic search enabled us to discover relevant content that may have been missed by a traditional keyword search. It facilitates the exploration of related ideas, synonyms, and contextually relevant information, leading to a more comprehensive and accurate retrieval of desired results. After calculating the cosine similarity between the vector representations of the paragraphs and keywords, we extracted those paragraphs that are related to the bioeconomy.

Figure 2 presents a density plot illustrating the distribution of cosine similarity values between the paragraphs and different exemplary keywords. To isolate scores indicating significant relevance, we computed the z-score corresponding to a two-tailed significance level of 0.01, yielding a critical value of approximately 2.576. This critical value helped delineate an upper threshold, represented by the red dashed lines on the histograms. Scores surpassing this threshold are considered statistically significant at the 0.01 level. Such high scores indicate texts with pronounced relevance to the selected keywords, marking them as candidates for in-depth examination. Following this filtering step, the dataset retained approximately 4.6 million paragraphs, constituting around 12.5 % of

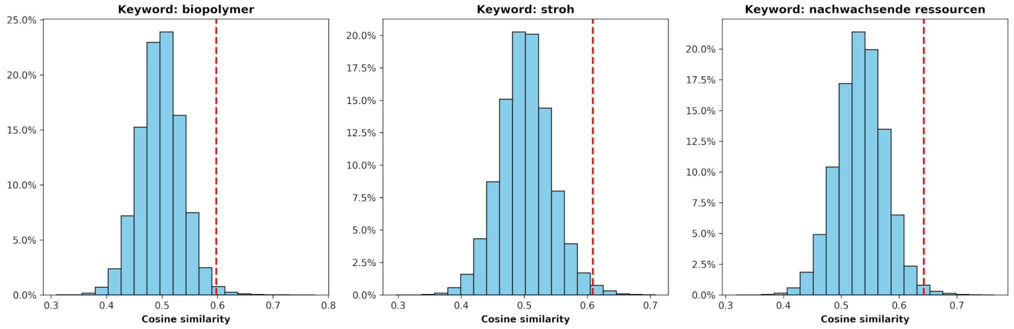


Figure 2: Distribution of cosine similarity values for selected keywords

the initial corpus.

In order to differentiate the technological capabilities of the extracted firms, we employ an advanced text classification approach. This approach builds upon the results obtained from semantic search and incorporates sophisticated techniques to accurately assess and distinguish the varying levels of technological capabilities exhibited by each firm. By leveraging this comprehensive text classification methodology, we can provide a more detailed and nuanced understanding of the technological landscape among the analyzed firms. To accomplish this, we extracted 1460 random paragraphs from the results of the semantic search. These paragraphs were then manually labeled to differentiate between three distinct levels of technological capabilities: (1) *No bioeconomy*, representing general information unrelated to the bioeconomy, (2) *bioeconomy in general*, encompassing sectors such as forestry and wood processing companies, (3) *bioeconomy high-tech*, comprising advanced and knowledge-intensive fields such as biotechnology and bio-pharmaceuticals. We provide a table with anchor examples from our annotation in Appendix C. This fine-grained labeling process allows for a detailed classification of the paragraphs, enabling us to assess effectively the technological capabilities of each firm within the bioeconomy domain.

For the training of the machine learning model, we again utilize a pre-trained SBERT model. Thanks to the inherent language understanding capabilities of those language models, we can fine-tune them with a relatively small number of manually annotated texts. By utilizing this transfer learning approach, we can efficiently adapt the models to our specific task of classifying technological capabilities in the bioeconomy domain, while minimizing the need for a large corpus of annotated data (Ruder et al. 2019). Given that a company's website typically consists of multiple subpages and paragraphs, it is plausible that different technological capabilities are assigned to a firm. To address this, we propose a straightforward heuristic where each company is labeled with the highest technological capability assigned to any of its relevant paragraphs. Consequently, we aim to provide a simplified but practical method for assigning labels to companies based on their most advanced technological capability as identified within their website content.

The dataset was divided into training, validation, and test sets to assess the performance of the model. After training the model on the training set and fine-tuning it using the validation set, we evaluated its accuracy on the test set. The achieved overall accuracy of 87.67 % indicates that the model was able to correctly classify the technological capabilities of the firms with a very high level of accuracy. The model has a precision of 88.68 % and a recall of 86.6 %, indicating its ability to accurately predict positive outcomes and capture true positives. The F1 score of 87.63 % demonstrates a well-balanced integration of precision and recall (Manning et al. 2008, Bishop 2006). By utilizing the knowledge and patterns learned during the training process, the model generated predictions for each paragraph, assigning them to one of the pre-defined classes representing different levels of technological capabilities within the bioeconomy domain.

Table 1: Description of the data set

Number of firms with a website	678,381
Number of bioeconomy firms (all)	142,949
Share of bioeconomy firms	21.07 %
Number of high-tech bioeconomy firms	13,554
Share of high-tech bioeconomy firms among all bioeconomy firms	9.48 %
Largest bioeconomy topics	Timber construction, agriculture, textiles
Average number of bioeconomy activities (topics) per firm	5.43

3.3 Uncovering different economic activities within the bioeconomy

While our previous efforts have focused on classifying bioeconomy firms based on their technological capabilities, there is an opportunity to delve deeper into the economic activities within the bioeconomy. This avenue of exploration can be effectively realized through the application of topic modeling techniques, which unveil underlying economic themes and patterns embedded within the dataset (Dahlke et al. 2024). To cluster the vector representations of the identified paragraphs, we utilize the BERTopic framework (Grootendorst 2022). Given the complexity of the vector embeddings, we implement a dimensionality reduction step. This process simplifies the vectors by distilling their information down to the most fundamental features. Such a simplification leads to more effective clustering, as it enhances the distinctiveness of the documents. We use UMAP for dimensionality reduction, which balances the preservation of essential local and global data structures (McInnes et al. 2018). Local data structures refer to the subtle relationships and patterns between neighboring data points, which are crucial for capturing the similarities between paragraphs. Global data structures, on the other hand, represent the broader distribution and relationships across the dataset, which are essential for preserving overarching thematic connections in the semantic space (McInnes et al. 2018). After dimensionality reduction, HDBSCAN is used for density-based clustering (McInnes et al. 2017). This technique effectively adapts to clusters of different shapes and densities, while effectively distinguishing between core topics and outliers. Finally, a modified version of TF-IDF, which focuses on clusters rather than documents, allows the identification of distinctive words that characterize each cluster.

The topic modeling methodology was extended to encompass all 1,066,059 paragraphs attributed to bioeconomy companies. In this pursuit, we chose a minimum cluster size of 1500, guided by the objective to construct clusters of meaningful coherence and comprehensive representation. In a concluding manual phase, we further clustered similar topics and subsequently refined the topic descriptions. In total, we extracted 55 topics from the corpus. For comprehensive insights into the results of our topic modeling analysis, please refer to Appendix E.

4 Results: Understanding the geography of bioeconomy firms

In total, we have identified 142,949 companies operating within the bioeconomy domain, of which 13,554 are classified as high-tech bioeconomy firms. Table 1 provides an overview of the dataset. Figure 3 shows the density distribution of bioeconomy firms in hexagonal cells.

4.1 The urban-rural divide in the bioeconomy

In order to gain insights into the geographical distribution of bioeconomy-related firms, we calculated the proportion of companies classified as either "bioeconomy" or "bioeconomy high-tech" within the entire cohort of observed web companies at the NUTS-3 level. Figure 4 shows the resulting map of firms active in the bioeconomy domain. We observe pronounced spatial disparities in the distribution of bioeconomy firms. In our overall findings, we observe a higher proportion of bioeconomy firms in regions characterized by lower population density. Regions with a high concentration of bioeconomy firms

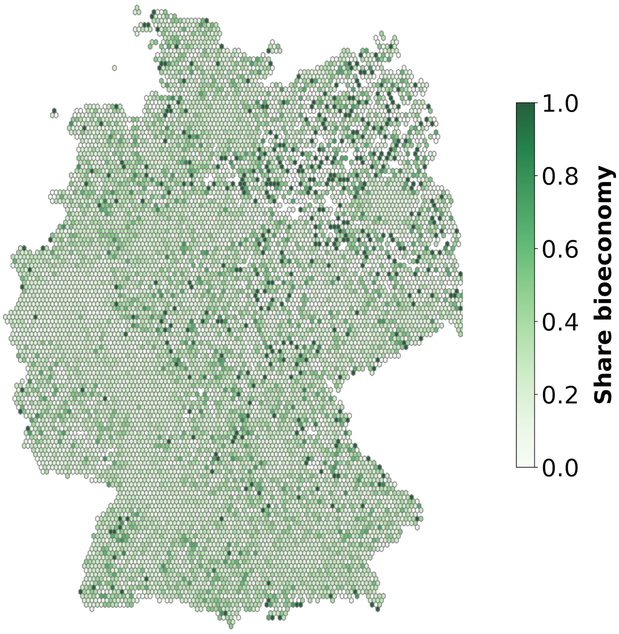


Figure 3: Share of bioeconomy firms

are, e.g., Südliche Weinstraße, Lüchow-Dannenberg, Oberallgäu, Regen or Cloppenburg (district).

Figure 4 also shows the bivariate relationship between the share of bioeconomy firms and population density. The inverse relationship observed between population density and bio-based firms aligns with the visual pattern depicted in the map. This negative correlation substantiates the initial visual impression, indicating a higher prevalence of bioeconomy firms in regions with lower population density.

To enhance the granularity of our findings, we classified the regions into distinct categories based on settlement types in Germany as defined by the BBSR (Federal Institute for Research on Building, Urban Affairs, and Spatial Development). Our results, as presented in Figure 4, reveal a discernible pattern that provides additional support for our earlier findings. Notably, urban centers have a smaller proportion of bioeconomy companies, while more thinly populated counties on average have a higher proportion of bioeconomy companies. To test the observed differences statistically, we ran a Welch-ANOVA, which confirmed that these differences are statistically significant ($p < 0.001$), with the exception of the comparison between “rural district with urbanization tendencies” and “sparsely populated rural district”. We employed a standard urban scaling framework (Bettencourt 2013, Broekel et al. 2023) as an alternative approach to scrutinize the geography of bio-based activities. Our findings reveal a scaling coefficient below one, signifying that bioeconomy firms are inclined to be situated in regions with a lower overall number of companies. Specifically, a 10 % surge in the number of companies within a given region corresponds to a 9.5 % increase in the number of bioeconomy firms, indicative of a sublinear scaling. This suggests that bioeconomy firms generally tend to thrive in less metropolitan areas. In summary, our results support the first hypothesis that bioeconomy firms are more likely to be found in rural regions.

4.2 The geography of high-tech bioeconomy firms

So far, we have shown the geography of bioeconomy companies in general. The following section discusses the geography of high-tech activities within the bioeconomy. Figure 5 shows the share of high-tech bioeconomic activities among all identified bioeconomic activities. Contrary to the general findings on the bioeconomy, high-tech activities are primarily concentrated in large cities. Districts with particularly large shares of high-tech activities are Jena, Heidelberg, Darmstadt and Munich. Moreover, Figure 5 provides an

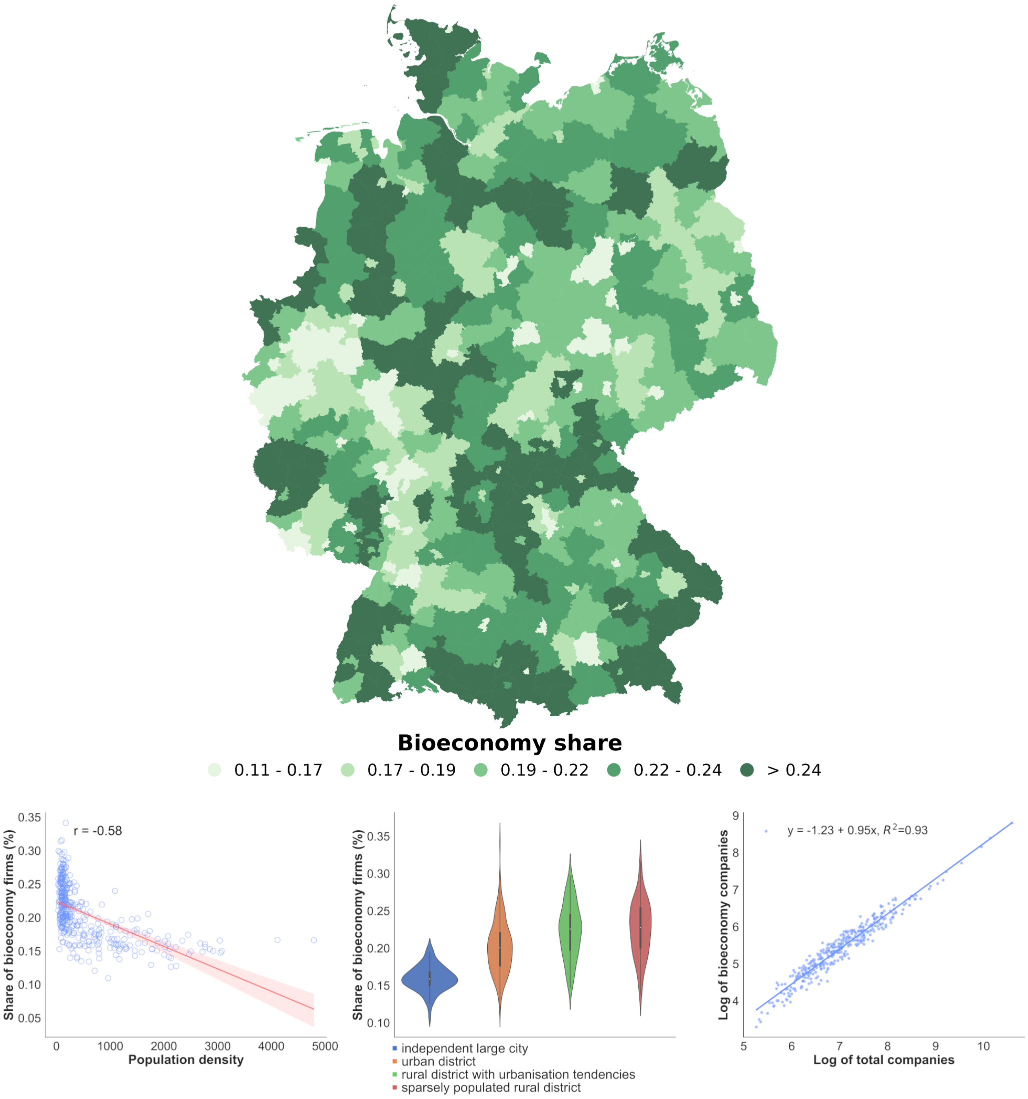


Figure 4: Geographical distribution of bioeconomy companies

indication of a positive linear correlation between the population density of a region and the share of high-tech bioeconomy firms. The variation in the share of high-tech bioeconomy firms between rural and urban regions is also well illustrated in Figure 5, which visualizes the distribution along settlement types. We find statistically significant ($p < 0.001$) differences between all pairwise comparisons except the comparison between "rural district with urbanization tendencies" and "sparsely populated rural district". The scaling analysis reinforces these findings, revealing a superlinear scaling coefficient of 1.24. This signifies that a 10 % augmentation in the number of firms within a region corresponds to a 12.4 % increase in the number of high-tech bioeconomy firms in that region. The scaling coefficients of high-tech bioeconomy firms are remarkably similar to those derived from analyses using patent data (Bettencourt 2013, Broekel et al. 2023). This further strengthens the robustness of our findings, suggesting a consistent pattern of growth dynamics across different indicators of technological innovation and economic development. This observation suggests that high-tech bioeconomy firms thrive in metropolitan regions, propelled by urbanization economies. In conclusion, our results seem to indicate that the second hypothesis, namely the spatial concentration of high-tech firms in urban regions, is also empirically supported, although the results are less clear than in the analysis of the first hypothesis. This is due to a number of rural regions specializing in high-tech activities, as can be seen in Figure 5.

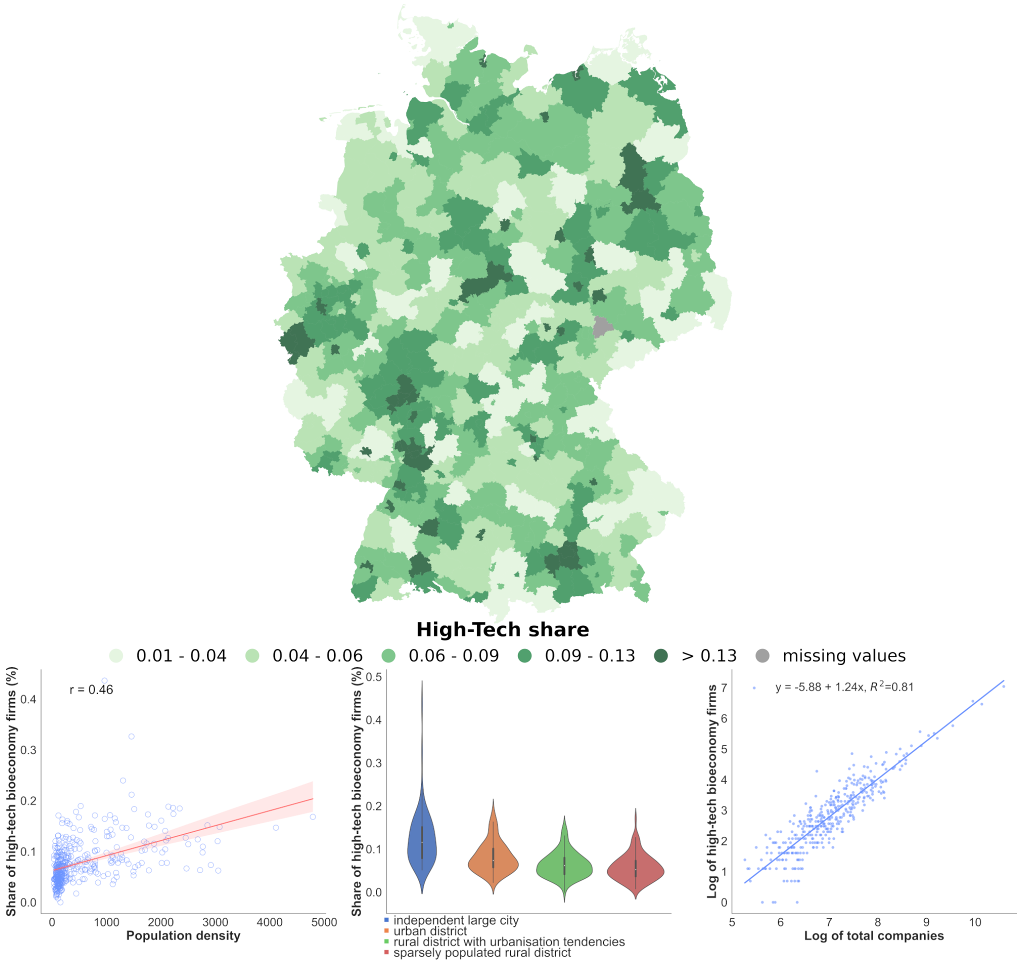


Figure 5: Geographical distribution of high-tech bioeconomy companies

4.3 The relationship between land use and economic activities in the bioeconomy

In the next step, we use the results of the topic modeling to gain further insights into the geography of bioeconomy firms in Germany. The topics identified provide information on the business segments in which a bioeconomy firm is active. The topics therefore go a step further than our delineation of high-tech companies that we used for the analysis of the second hypothesis. The topics should not be regarded as being directly equivalent to industry classifications, technology classes or economic sectors, but they do provide an indication of the business segments in which a company is active. The topics also provide an indication of which biogenic resources a firm uses for its business activities. We use this information to assess the third hypothesis, namely the spatial proximity of bioeconomy firms to their biomass needs. Figure 6 presents the correlation of the shares of different topics of all activities of the bioeconomy firms in a region with the share of the region's area that can be associated as a natural resource to a given topic. For example, one scatterplot illustrates the significant positive correlation of the activities of local bioeconomy companies related to “Wood” and the forest area of a region. The higher the share of forest area in a region, the higher the share of firms whose business activities are related to wood.

In conclusion, our analysis reveals that numerous hypotheses, which regional researchers might presuppose as applicable to the geography of bioeconomy firms, are indeed corroborated by the dataset we have compiled. We extended our analysis, as presented here, to the granularity of labor market regions, reinforcing the identified hypotheses. A corresponding set of figures for labor market regions is included in Appendix

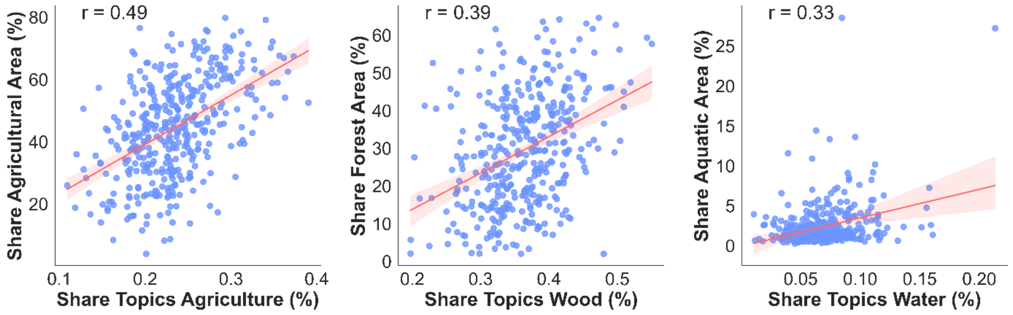


Figure 6: Land use and economic activities

D. As the main contribution of our work, we have presented a novel dataset, which will be a valuable resource for researchers focusing on the geography of bioeconomy activities. We have made an aggregated version of this dataset accessible in the online appendix. For each NUTS-3 region in Germany, the dataset encompasses several key variables, such as the total number of companies with a website, the count of firms identified as part of the bioeconomy, and metrics related to the various bioeconomy activities we have categorized. A comprehensive description of all variables is also provided in the online appendix. In utilizing the data generated in this paper, we urge researchers to contemplate the potential pitfalls associated with novel (web) data sources in regional research - a consideration familiar to many quantitative social scientists and economists (Einav, Levin 2014, Franklin 2022, 2023, Kitchin 2013).

The dataset can be utilized in diverse ways for quantitative research, such as linking bioeconomy activities with regional variables to understand the drivers of regional bioeconomy activities, following the increasing research on regional determinants of green economic activities (Losacker et al. 2023a). Alternatively, the data could serve as an independent variable to explore regional impacts, like the effect of local bioeconomy activities on regional development (e.g., value added, employment) or their correlation with environmental indicators (e.g., emissions, biodiversity loss). Qualitatively, the dataset can help identify regions for in-depth case studies, including those with notable bioeconomy activity or those with limited bioeconomy involvement.

5 Conclusion

This paper's main objective was to craft a methodological approach for the comprehensive measurement of bio-based economic activities, with an added focus on revealing their geographical distribution. This research goal was driven by the realization that traditional statistical classifications, be it industry or technology categories, fall short in encapsulating the nuances of bioeconomic activities. Such limitations not only deprive researchers of robust data for bioeconomy analysis, but also impede policymakers in their pursuit of evidence-informed decisions.

Against this background, we have built a unique dataset that enables us to identify and map bioeconomy firms in Germany. The dataset is based on a web-mining approach, using the open-source web repository CommonCrawl to identify German company websites. From this data, we have identified bioeconomy firms using a combination of different natural language processing techniques, utilizing the semantic capabilities of modern transformer models. Our final dataset enables a precise analysis of the bioeconomy, its geography, and its various domains. We used this dataset to test three hypotheses about the bioeconomy, thereby assessing the applicability of the dataset and demonstrating its potential for empirical research. First, we showed that bioeconomy firms predominantly concentrate in rural areas. Second, however, we demonstrated that high-tech activities related to the bioeconomy concentrate in urban areas. Third, we found that economic activities centered on bio-based processes and biomass locate in close proximity to their primary biomass feedstocks.

However, it is important to interpret these results with caution and keep in mind several limitations. The presence of a firm's website varies significantly based on specific firm attributes. Consequently, our suggested web mining framework may not be applicable for analyzing certain firms. In particular, firms that are either very new or very small, along with those operating in specific sectors and regions, tend to have limited website availability (see Kinne, Axenbeck 2020). For our study, this implies a particular gap in information regarding small agricultural firms located in rural areas, which are underrepresented due to these constraints. Furthermore, webpage data is subject to a self-description bias, since firms have the autonomy to choose both the nature and the manner in which information is presented on their sites.

We provide the compiled dataset in an aggregated form along with variable descriptions in the online appendix. We encourage fellow researchers to utilize this dataset to address the numerous unresolved research questions surrounding the bioeconomy. We are confident that future analyses of this data will yield important insights, paving the way for place-specific recommendations that can inform industrial and innovation policies geared towards sustainable regional development.

Acknowledgements

We acknowledge financial support by the German Federal Ministry of Education and Research (BMBF 031B1281). Previous versions of this article were presented at the 16th Rauischholzhausen symposium on Economic Geography (April 2023), at the RSA Annual Conference in Ljubljana (June 2023), at an annual meeting of junior bioeconomy research groups funded by the BMBF (November 2023), and at the 7th Geography of Innovation Conference in Manchester (January 2024). The authors have benefited greatly from feedback received from the participants of these events. The methodological approach employed in this paper has been discussed during a workshop on advanced methods for the measurement of economic and innovation activities in the bioeconomy, held at Giessen University in August 2023. We thank the discussants, in particular Henning Kroll, Daniel Schiller and Sven Wydra, for feedback during this workshop.

Declaration of interest

No potential competing interest was reported by the authors.

References

Abbasiharofteh M, Broekel T (2020) Still in the shadow of the wall? The case of the Berlin biotechnology cluster. *Environment and Planning A: Economy and Space*: 0308518X2093390. CrossRef

Abbasiharofteh M, Krüger M, Kinne J, Lenz D, Resch B (2023) The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis* 18: 507–529. CrossRef

Aguilar A, Wohlgemuth R, Twardowski T (2018) Perspectives on bioeconomy. *New Biotechnology* 40: 181–184. CrossRef

Allain S, Ruault JF, Moraine M, Madelrieux S (2022) The 'bioeconomics vs bioeconomy' debate: Beyond criticism, advancing research fronts. *Environmental Innovation and Societal Transitions* 42: 58–73. CrossRef

Andersson I, Grundel I (2021) Regional policy mobilities: Shaping and reshaping bioeconomy policies in Värmland and Västerbotten, Sweden. *Geoforum* 121: 142–151. CrossRef

Asheim B, Grillitsch M, Trippl M (2016) Regional innovation systems: Past – presence – future. In: Shearmur R, Carrincazeaux C, Doloreux D (eds), *Handbook on the geographies of innovation.* Edward Elgar Publishing Ltd., 45–62. CrossRef

Bauer F (2018) Narratives of biorefinery innovation for the bioeconomy: Conflict, consensus or confusion? *Environmental Innovation and Societal Transitions* 28: 96–107. CrossRef

Bauer F, Hansen T, Hellsmark H (2018) Innovation in the bioeconomy–dynamics of biorefinery innovation networks. *Technology Analysis and Strategic Management* 30: 935–947. CrossRef

Befort N (2023) *The Bioeconomy: Institutions, Innovation and Sustainability*. Routledge. CrossRef

Bettencourt LM (2013) The origins of scaling in cities. *Science* 340: 1438–1441. CrossRef

Bishop C (2006) *Pattern Recognition and Machine Learning*. Springer New York. CrossRef

Bringezu S, Distelkamp M, Lutz C, Wimmer F, Schaldach R, Hennenberg KJ, Böttcher H, Egenolf V (2021) Environmental and socioeconomic footprints of the German bioeconomy. *Nature Sustainability* 4: 775–783. CrossRef

Broekel T, Knuepling L, Mewes L (2023) Boosting, sorting and complexity – urban scaling of innovation around the world. *Journal of Economic Geography*. CrossRef

Bugge M, Hansen T, Klitkou A (2016) What is the bioeconomy? A review of the literature. *Sustainability* 8: 691. CrossRef

Cooke P (2002) Biotechnology clusters as regional, sectoral innovation systems. *International Regional Science Review* 25: 8–37. CrossRef

Dahlke J, Beck M, Kinne J, Lenz D, Dehghan R, Wörter M, Ebersberger B (2024) Epidemic effects in the diffusion of emerging digital technologies: Evidence from artificial intelligence adoption. *Research Policy* 53: 104917. CrossRef

Ehrenfeld W, Kropfhäußer F (2017) Plant-based bioeconomy in Central Germany – A mapping of actors, industries and places. *Technology Analysis & Strategic Management* 29: 514–527. CrossRef

Einav L, Levin J (2014) Economics in the age of big data. *Science* 346: 6210. CrossRef

El-Chichakli B, von Braun J, Lang C, Barben D, Philp J (2016) Policy: Five cornerstones of a global bioeconomy. *Nature* 535: 221–223. CrossRef

Fischer L, Losacker S, Wydra S (2024) National specialization and diversification in the bioeconomy: Insights from biobased technologies in chemical and pharmaceutical sectors. *Technology in Society* 76: 102462. CrossRef

Franklin R (2022) Quantitative methods I: Reckoning with uncertainty. *Progress in Human Geography* 46: 689–697. CrossRef

Franklin R (2023) Quantitative methods II: Big theory. *Progress in Human Geography* 47: 178–186. CrossRef

Friedrich J, Bunker I, Uthes S, Zscheischler J (2021) The potential of bioeconomic innovations to contribute to a social-ecological transformation: A case study in the livestock system. *Journal of Agricultural and Environmental Ethics* 34: 1–26. CrossRef

Giurca A, Befort N (2023) Deconstructing substitution narratives: The case of bioeconomy innovations from the forest-based sector. *Ecological Economics* 207: 107753. CrossRef

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794v1

Gök A, Waterworth A, Shapira P (2015) Use of web mining in studying innovation. *Scientometrics* 102: 653–671. CrossRef

Haarich S, Kirchmayr-Novak S European Commission, Joint Research Centre, Publications Office of the European Union, Luxembourg. CrossRef

Halonen M, Näyhä A, Kuhmonen I (2022) Regional sustainability transition through forest-based bioeconomy? Development actors' perspectives on related policies, power, and justice. *Forest Policy and Economics* 142: 102775. CrossRef

Heimeriks G, Boschma R (2014) The path- and place-dependent nature of scientific knowledge production in biotech 1986-2008. *Journal of Economic Geography* 14: 339–364. CrossRef

Hermans F (2018) The potential contribution of transition theory to the analysis of bioclusters and their role in the transition to a bioeconomy. *Biofuels, Bioproducts and Biorefining* 12: 265–276. CrossRef

Hermans F (2021) Bioclusters and sustainable regional development. In: Sedita SR, Blasi S (eds), *Rethinking Clusters: Place-based Value Creation in Sustainability Transitions*. Springer, Cham, 81–91. CrossRef

Imbert E, Ladu L, Morone P, Quitzow R (2017) Comparing policy strategies for a transition to a bioeconomy in Europe: The case of Italy and Germany. *Energy Research & Social Science* 33: 70–81. CrossRef

Jander W, Grundmann P (2019) Monitoring the transition towards a bioeconomy: A general framework and a specific indicator. *Journal of Cleaner Production* 236. CrossRef

Kamath R, Elola A, Hermans F (2023) The green-restructuring of clusters: Investigating a biocluster's transition using a complex adaptive system model. *European Planning Studies* 31: 1842–1867. CrossRef

Kinne J, Axenbeck J (2020) Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics* 125: 2011–2041. CrossRef

Kinne J, Lenz D (2021) Predicting innovative firms using web mining and deep learning. *PLOS ONE* 16: e0249071. CrossRef

Kitchin R (2013) Big data and human geography. *Dialogues in Human Geography* 3: 262–267. CrossRef

Kriesch L (2023) *Web Mining und Natural Language Processing als methodisches Komplement in der Wirtschaftsgeographie*. CrossRef

Kuckertz A, Berger ES, Brändle L (2020) Entrepreneurship and the sustainable bioeconomy transformation. *Environmental Innovation and Societal Transitions* 37: 332–344. CrossRef

Laasonen V (2023) Building dynamic capabilities in the transition toward a knowledge-based bioeconomy: A case study of three Finnish regions. *Regional Studies*: 1–12. CrossRef

Lasarte Lopez J, González Hermoso H, Rossi Cervi W, Van Leeuwen M, M'barek R (2023) BioRegEU. a pilot dataset for regional employment and value added in the EU bioeconomy. Publications office of the european union. CrossRef

Losacker S, Hansmeier H, Horbach J, Liefner I (2023a) The geography of environmental innovation: A critical review and agenda for future research. *Review of Regional Research*: 1–26. CrossRef

Losacker S, Heiden S, Liefner I, Lucas H (2023b) Rethinking bioeconomy innovation in sustainability transitions. *Technology in Society* 74: 102291. CrossRef

Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval*. Cambridge University Press. CrossRef

Martin H, Coenen L (2014) Institutional context and cluster emergence: The biogas industry in Southern Sweden. *European Planning Studies* 23: 2009–2027. CrossRef

Martin H, Grundel I, Dahlström M (2023) Reconsidering actor roles in regional innovation systems: Transformative industrial change in the forest-based bioeconomy. *Regional Studies* 57: 1636–1648. CrossRef

McInnes L, Healy J, Astels S (2017) hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2: 205. CrossRef

McInnes L, Healy J, Melville J (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *The Journal of Open Source Software* 3: 861. CrossRef

Morales D, Dahlström M (2022) Smart specialization and participatory processes in green path renewal. Analysis of the forest-based bioeconomy in sparsely populated regions in the Nordics. *European Planning Studies*: 1–20. CrossRef

Ozgun B, Broekel T (2022) Assessing press releases as a data source for spatial research. *REGION* 9: 25–44. CrossRef

Patermann C, Aguilar A (2021) A bioeconomy for the next decade. *EFB Bioeconomy Journal* 1: 100005. CrossRef

Prochaska L, Schiller D (2021) An evolutionary perspective on the emergence and implementation of mission-oriented innovation policy: The example of the change of the leitmotif from biotechnology to bioeconomy. *Review of Evolutionary Political Economy* 2: 141–249. CrossRef

Prochaska L, Schiller D (2024) Spatial distribution of bioeconomy R&D funding: Opportunities for rural and lagging regions? *European Planning Studies*: 1–21. CrossRef

Proestou M, Schulz N, Feindt PH (2023) A global analysis of bioeconomy visions in governmental bioeconomy strategies. *Ambio 2023*: 1–13. CrossRef

Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, Aslanides J, Henderson S, Ring R, Young S, Rutherford E, Hennigan T, Menick J, Cassirer A, Powell R, Driessche GVD, Hendricks LA, Rauh M, Huang PS, Glaese A, Welbl J, Dathathri S, Huang S, Uesato J, Mellor J, Higgins I, Creswell A, Mcaleese N, Wu A, Elsen E, Jayakumar S, Buchatskaya E, Budden D, Sutherland E, Simonyan K, Paganini M, Sifre L, Martens L, Li L, Kuncoro A, Nematzadeh A, Gribovskaya E, Donato D, Lazaridou A, Mensch A, Lespiau JB, Tsimpoukelli M, Grigorev N, Fritz D, Sottiaux T, Pajarskas M, Pohlen T, Gong Z, Toyama D, D'autume CDM, Li Y, Terzi T, Mikulik V, Babuschkin I, Clark A, De D, Casas L, Guy A, Jones C, Bradbury J, Johnson M, Hechtman B, Weidinger L, Gabriel I, Isaac W, Lockhart E, Osindero S, Rimell L, Dyer C, Vinyals O, Ayoub K, Stanway J, Bennett L, Hassabis D, Kavukcuoglu K, Irving G (2021) Scaling language models: Methods, analysis & insights from training Gopher. https://arxiv.org/abs/2112.11446v2

Ramirez P (2021) Technological revolutions, socio-technical transitions and the role of agency: Värmland's transition to a regional bio-economy. *Regional Studies* 55: 1642–1651. CrossRef

Refsgaard K, Kull M, Slätmo E, Meijer MW (2021) Bioeconomy – A driver for regional development in the Nordic countries. *New Biotechnology* 60: 130–137. CrossRef

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. https://github.com/UKPLab/

Ronzon T, Piotrowski S, M'Barek R, Carus M (2017) A systematic approach to understanding and quantifying the EU's bioeconomy. *Bio-based and Applied Economics* 6: 1–17. CrossRef

Ruder S, Peters ME, Swayamdipta S, Wolf T (2019) Transfer learning in natural language processing. *Proceedings of the 2019 Conference of the North*: 15–18. CrossRef

Sanz-Hernández A, Sanagustín-Fons MV, López-Rodríguez ME (2019) A transition to an innovative and inclusive bioeconomy in Aragon, Spain. *Environmental Innovation and Societal Transitions* 33: 301–316. CrossRef

Steinböck N, Trippl M (2023) The thorny road towards green path development: The case of bioplastics in Lower Austria. *Regional Studies, Regional Science* 10: 735–749. CrossRef

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Łukasz Kaiser, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vogelpohl T, Töller AE (2021) Perspectives on the bioeconomy as an emerging policy field. *Journal of Environmental Policy & Planning* 23: 143–151. CrossRef

Wesseler J, von Braun J (2017) Measuring the bioeconomy: Economics and policies. *Annual Review of Resource Economics* 9: 275–298. CrossRef

Wydra S (2020) Measuring innovation in the bioeconomy — Conceptual discussion and empirical experiences. *Technology in Society* 61: 101242. CrossRef

A Appendix: Text quality filtering

We adapted the text quality filtering heuristics established by Rae et al. (2021) to better suit the nuances of the German language, implementing the following modifications to the filtering process:

- Paragraphs are excluded if the average word length falls outside the 3 to 12 character range.
- Paragraphs are eliminated if the ratio of symbols to words exceeds 0.15.
- Paragraphs are discarded if they contain fewer than two stopwords from either English or German.
- Any paragraph composed entirely of uppercase letters is also removed from consideration, as this often signifies non-standard text or spam.

B Appendix: Keyword list

Table B.1: Keyword list

biobasiert	Biokraftstoff	Naturfasern	Agrar	Bioökonomie
Holz	Biotenside	Biopharmazeutika	Aquakulturen	Landwirtschaft
Zellstoff	Bioschmierstoff	Mikroorganismen	Mikrobiom	Aquaponik
Biophysik	Bioengineering	Enzym	Biolösungsmittel	Peptide
Biotechnik	Biomasse	Biokunststoff	Biopolymer	Nukleoside
Biotechnologie	nachwachsende Rohstoffe	Bioplastik	biologisch abbaubar	Vertikale Landwirtschaft
Bioenergie	biogen	Pflanze	Papier	Nachwachsende Ressourcen
Biochemie	biologisch	Forst	Fischerei	Bioklebstoff
Biokosmetik	pflanzenbasiert			

C Appendix: Anchor examples of annotated training data

Table C.2: Anchor examples of annotated training data

Label	English (translated)	German (original)
Bioeconomy general	HOLZ-BARAN is your partner for woodworking, floor coverings and solid wood flooring in Luppa near Leipzig in Saxony.	HOLZ-BARAN ist Ihr Partner rund um Holzbearbeitung, Bodenbeläge und Massivholzdielen in Luppa bei Leipzig in Sachsen.
Bioeconomy general	Raiffeisen Bio-Brennstoffe GmbH sells wood chips, biomass, wood briquettes and, in particular, wood pellets. Together with co-operative partners, the associated company of AGRAVIS Raiffeisen AG has built up an efficient sales network throughout the northern half of Germany.	Die Raiffeisen Bio-Brennstoffe GmbH vertreibt Hackschnitzeln, Biomasse, Holzbriketts und insbesondere Holzpellets. Gemeinsam mit genossenschaftlichen Partnern hat das Beteiligungsunternehmen der AGRAVIS Raiffeisen AG ein leistungsfähiges Vertriebsnetz in der gesamten Nordhälfte Deutschlands aufgebaut.
Bioeconomy general	Our meat factory consists of several organic farms, all of which are located in our beautiful district of Höxter in East Westphalia. In the easternmost corner of NRW, where there are still many meadows and pastures!	Unsere Fleischmanufaktur besteht aus mehrere Bio-Höfen, die alle in unserem schönem Kreis Höxter in Ostwestfalen angesiedelt sind. Im östlichsten Fleck von NRW, wo es noch viele Wiesen & Weiden gibt!

Table C.2: Anchor examples of annotated training data - continued

Label	English (translated)	German (original)
Bioeconomy high-tech	We are BioCer Entwicklungs-GmbH from Bayreuth. As a young, innovative medical technology company, we specialise in the research, development and production of innovative medical products made from biomaterials without animal or human components. As a service provider, we also coat implants for our customers and develop new types of medical products.	Wir sind die BioCer Entwicklungs-GmbH aus Bayreuth. Als junges, innovatives Unternehmen der Medizintechnik haben wir uns auf die Forschung, Entwicklung und Produktion von innovativen Medizinprodukten aus Biomaterialien ohne tierische oder humane Bestandteile spezialisiert. Darüber hinaus beschichten wir als Dienstleister für unsere Kunden Implantate und entwickeln neuartige Medizinprodukte.
Bioeconomy high-tech	Hansen is a global biotechnology company that develops natural solutions for the food, nutrition, pharmaceutical and agricultural industries.	Hansen ist ein globales Biotechnologieunternehmen, das natürliche Lösungen für die Lebensmittel-, Ernährungs-, Pharma- und Landwirtschaftsindustrie entwickelt.
Bioeconomy high-tech	The BMW Group is developing innovative, bio-based surfaces in cooperation with start-up companies. For example, the newly developed DeserttexTM is made from powdered cactus fibres and a bio-based polyurethane matrix. This allows the elimination of animal-based raw materials to be combined with a reduction in CO2 emissions.	Die BMW Group entwickelt in Kooperation mit Start-up-Unternehmen innovative, biobasierte Oberflächen. So setzt sich z. B. das neuentwickelte DeserttexTM aus pulverisierten Kaktusfasern und einer biobasierten Polyurethan-Matrix zusammen. So lässt sich der Verzicht auf tierische Rohstoffe mit einer Co2-Reduzierung kombinieren.
No bioeconomy	Passport photos are subject to strict regulations (biometric photos). We are always familiar with the latest standards in order to be able to offer you successful passport photos at all times.	Passbilder unterliegen strengen Vorschriften (Biometrische Fotos). Wir sind immer mit den neuesten Standards vertraut, um Ihnen jederzeit gelungene Passbilder bieten zu können.
No bioeconomy	Today, slate is quarried underground and above ground using innovative and technically advanced processing methods in an environmentally friendly way with the aid of state-of-the-art technology. Efficient laying techniques make this sustainable natural product extremely economical.	Schiefer wird heute durch innovative und technisch weiterentwickelte Bearbeitungsmethoden umweltschonend mit Hilfe modernster Technik unter und über Tage abgebaut. Rationelle Verlegetechniken machen das nachhaltige Naturprodukt äußerst wirtschaftlich.

D Appendix: Findings for labor market regions

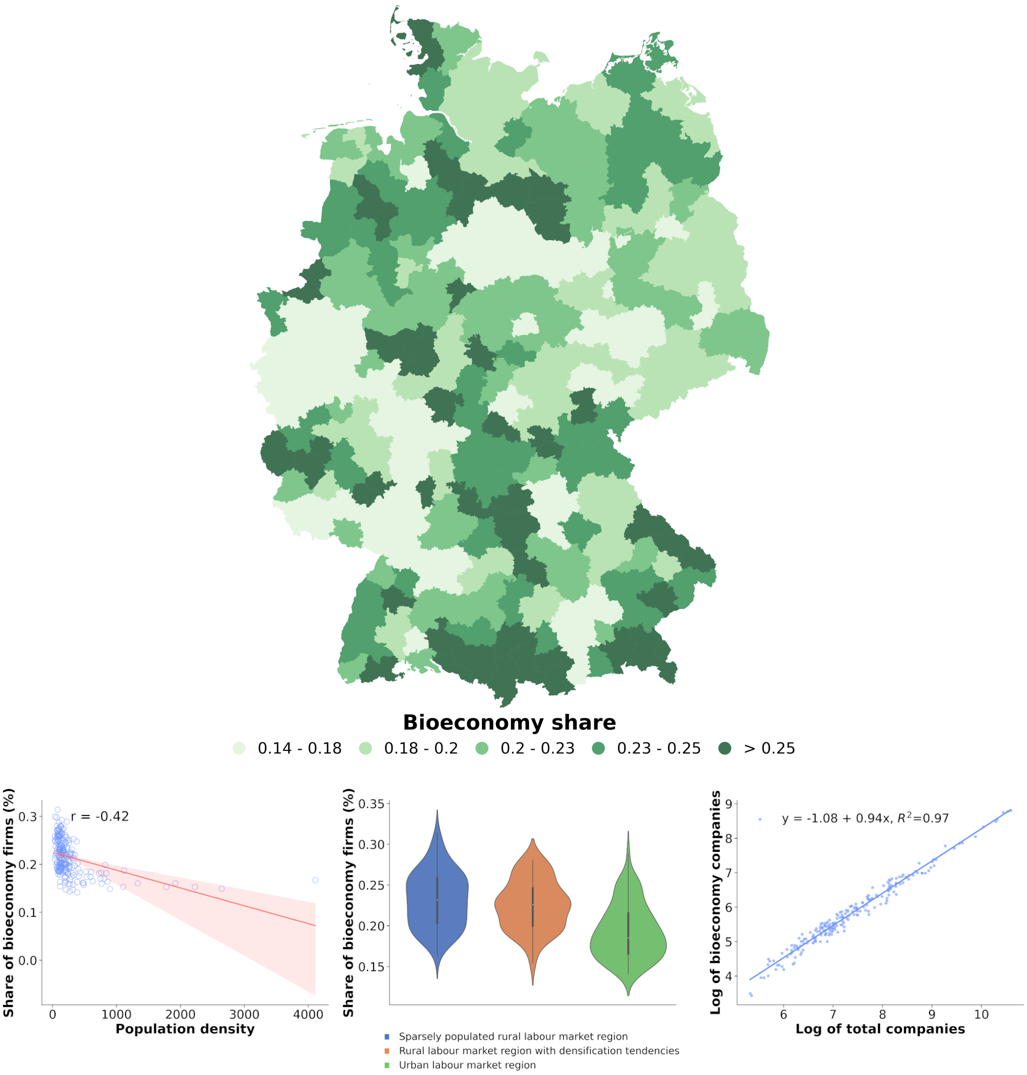

Bioeconomy share
0.14 - 0.18
0.18 - 0.2
0.2 - 0.23
0.23 - 0.25
> 0.25
Share of bioeconomy firms (%)
r = -0.42
0.3
0.2
0.1
0.0
0
1000
2000
3000
4000
Population density
Share of bioeconomy firms (%)
0.35
0.30
0.25
0.20
0.15
Sparsely populated rural labour market region
Rural labour market region with densification tendencies
Urban labour market region
Log of bioeconomy companies
y = -1.08 + 0.94x, R^2=0.97
9
8
7
6
5
4
6
7
8
9
10
Log of total companies

Figure D.1: Geographical distribution of bioeconomy companies (adapted version of Figure 4)

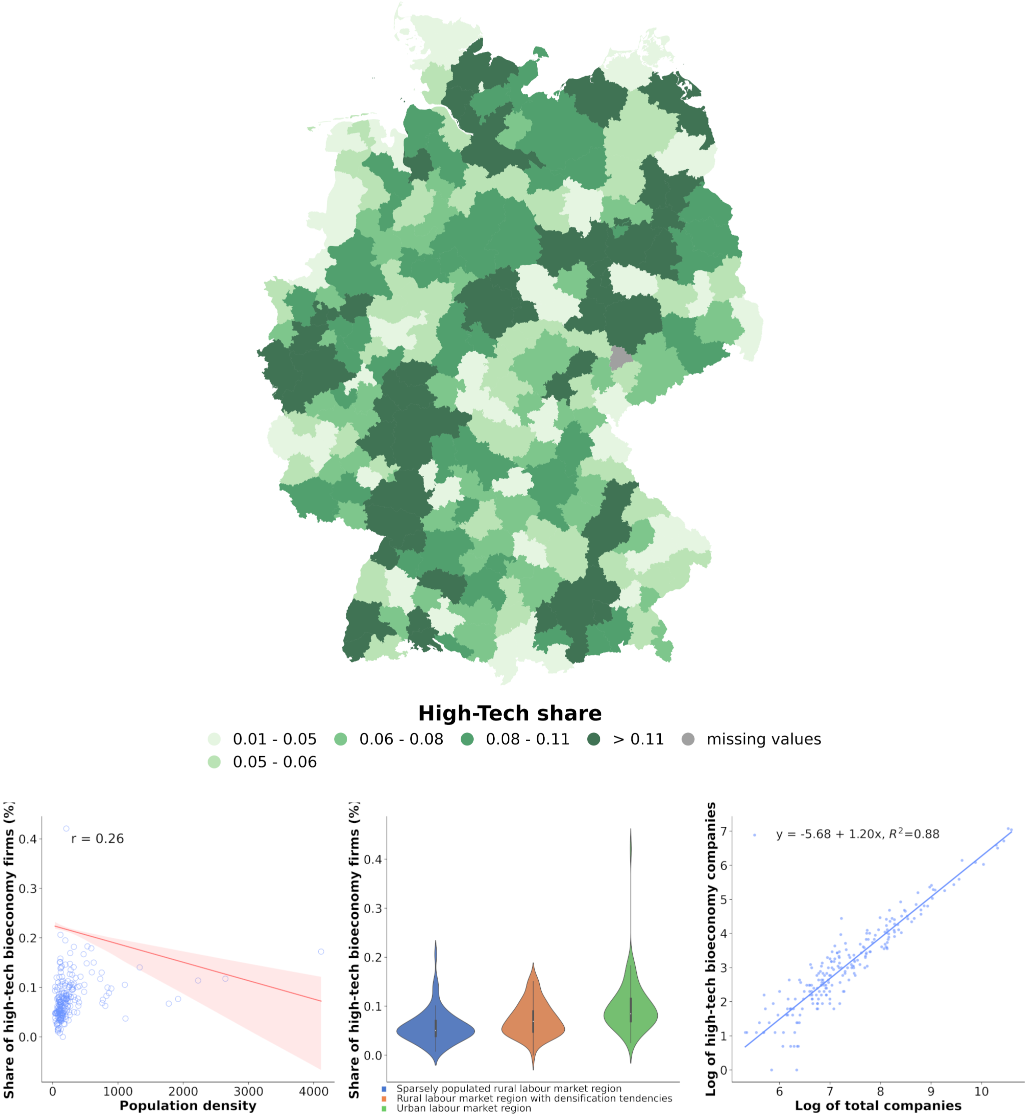

High-Tech share
0.01 - 0.05
0.05 - 0.06
0.06 - 0.08
0.08 - 0.11
> 0.11
missing values
Share of high-tech bioeconomy firms (%)
r = 0.26
0.4
0.3
0.2
0.1
0.0
0
1000
2000
3000
4000
Population density
Share of high-tech bioeconomy firms (%)
0.4
0.3
0.2
0.1
0.0
Sparsely populated rural labour market region
Rural labour market region with densification tendencies
Urban labour market region
Log of high-tech bioeconomy companies
y = -5.68 + 1.20x, R^2=0.88
7
6
5
4
3
2
1
0
6
7
8
9
10
Log of total companies

Figure D.2: Geographical distribution of bioeconomy companies (adapted version of Figure 5)

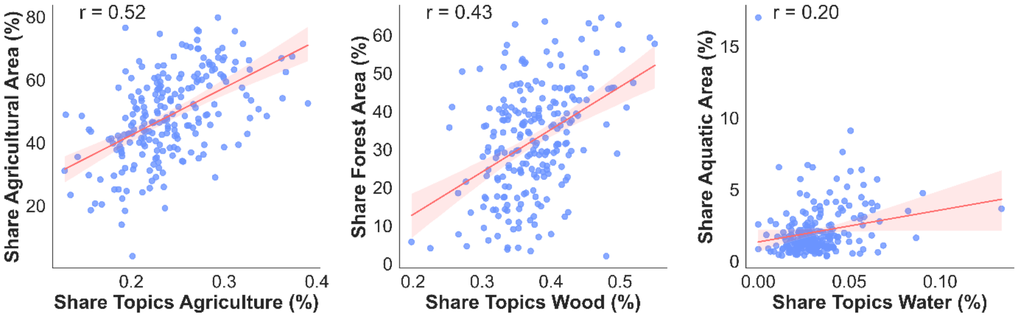


Figure D.3: Land use and economic activities (adapted version of Figure 6)

E Appendix: Topic model based on bioeconomy paragraphs

Table E.1: Topic model based on bioeconomy paragraphs

Topic	Name	Top words	Share
1	wood	holz', 'möbel', 'holzbau', 'bauen', 'materialien', 'haus', 'massivholz'	10.0 %
2	agriculture	bio', 'produkte', 'gemüse', 'landwirtschaft', 'obst', 'lebensmittel', 'qualität', 'region', 'ernährung'	8.5 %
3	miscellaneous	'products', 'research', 'food', 'high', 'well', 'quality', 'development', 'new', 'use', 'based'	7.6 %
4	textiles	baumwolle', 'leder', 'wolle', 'materialien', '100', 'material', 'farben', 'cm', 'teppiche'	6.3 %
5	agriculture	pflanzen', 'garten', 'blumen', 'blüten', 'rosen', 'stauden', 'pflege', 'pflanze', 'balkon'	5.2 %
6	wood	wald', 'bäume', 'wälder', 'natur', 'baum', 'mehr', 'co2', 'flächen'	4.9 %
7	cosmetics & food supplements	'haut', 'wirkung', 'inhaltsstoffe', 'haar', 'duft', 'aloe', 'vera', 'wirkt', 'naturkosmetik', 'öle'	3.8 %
8	food	fleisch', 'wurst', 'geschmack', 'steak', 'qualität', 'wurstwaren', 'rind', 'metzgerei', 'rindfleisch'	2.8 %
9	biotechnology	entwicklung', 'forschung', 'unternehmen', 'bereich', 'biotechnologie', 'dr', 'universität', 'entwickelt'	2.6 %
10	wood	holz', 'cm', 'kinder', 'gefertigt', 'material','spielzeug'	2.5 %
11	pulp and paper	papier', 'fsc', 'wellpappe', 'verpackungen', 'verpackung', 'rohstoffen', '100', 'biologisch', 'nachwachsenden'	2.3 %
12	animal feed	'futter', 'hunde', 'pferde', 'hund', 'pferd', 'fütterung', 'ernährung', 'vitamine'	2.2 %
13	wood	'pellets', 'holzpellets', 'heizen', 'holz', 'brennholz', 'brennstoff'	2.0 %
14	wood	parkett', 'laminat', 'bodenbelag', 'boden', 'parkettboden', 'dielen', 'böden', 'parkettböden'	2.0 %
15	food	wein', 'weine', 'trauben', 'weingut', 'reben', 'winzer', 'rebsorten', 'weinbau', 'weinberg'	2.0 %
16	agriculture	boden', 'dünger', 'pflanzen', 'nährstoffe', 'kompost', 'pflanze', 'düngung', 'erde', 'stickstoff', 'wachstum'	2.0 %
17	agriculture	tiere', 'hof', 'stall', 'kühe', 'rinder', 'schweine', 'betrieb', 'hühner', 'haltung', 'eier'	1.9 %
18	agriculture	biogas', 'biogasanlage', 'biomasse', 'biogasanlagen', 'strom', 'anlage', 'energie', 'anlagen', 'wärme', 'energien'	1.9 %
19	agriculture	früchte', 'sorten', 'ernte', 'erdbeeren', 'äpfel', 'sorte', 'obst', 'streuobstwiesen', 'apfel', 'beeren'	1.7 %
20	food	pfeffer', 'salz', 'tomaten', 'gemüse', 'salat', 'knoblauch', 'sauce', 'schneiden', 'olivenöl'	1.6 %
21	wood	wpc', 'holz', 'terrassendielen', 'terrasse', 'lärche', 'garten', 'gartenhaus', 'dielen', 'carport', 'sichtschutz'	1.5 %
22	wood	holz', 'oberfläche', 'holzes', 'pflege', 'außenbereich', 'schutz', 'oberflächen', 'holzschutz', 'öl', 'reinigung'	1.5 %
23	food	nüsse', 'zucker', 'zutaten', 'müsli', 'geschmack', 'lecker', 'einfach', 'kannst', 'mandeln', 'snack'	1.4 %
24	food	bienen', 'honig', 'insekten', 'imker', 'wildbienen', 'imkerei', 'manuka', 'nektar', 'bienenvölker', 'pollen'	1.4 %
25	agriculture	brot', 'getreide', 'mehl', 'weizen', 'dinkel', 'backen', 'backwaren', 'sauerteig', 'roggen', 'mühle'	1. 3%
26	wood	baum', 'bäume', 'baumpflege', 'bäumen', 'baumes', 'fällen', 'äste', 'baumfällung', 'fällung'	1.2%
27	wood	holz', 'fenster', 'türen', 'holzfenster', 'haustüren', 'alu', 'tür', 'kunststoff', 'innentüren'	1.1 %
28	water	fisch', 'lachs', 'fische', 'kaviar', 'forellen', 'fleisch', 'aquakultur', 'aal', 'matjes', 'forelle'	1.1 %
29	water	wasser', 'pflanzen', 'pflanze', 'erde', 'bewässerung', 'wurzeln', 'gießen', 'boden', 'blätter', 'topf'	1.1 %
30	food	kräuter', 'kräutern', 'wildkräuter', 'natur', 'pflanzen', 'küche', 'gewürze', 'gemüse', 'wildpflanzen', 'heilpflanzen'	1.0 %
31	food	käse', 'milch', 'käsesorten', 'molkerei', 'geschmack', 'käserei', 'rohmilch', 'joghurt', 'weichkäse'	0.9 %
32	agriculture	maschinen', 'ernte', 'bodenbearbeitung', 'mais', 'landwirtschaftlichen', 'einsatz', 'mähdrescher', 'technik', 'aussaat', 'landtechnik'	0.9 %

continued on the next page

Table E.1: Topic model based on bioeconomy paragraphs – continued

Topic	Name	Top words	Share
33	food	gin', 'rum', 'whisky', 'geschmack', 'aromen', 'aroma', 'botanicals', 'alkohol', 'vanille', 'destilliert'	0.8 %
34	food	kaffee', 'bohnen', 'kaffees', 'kaffeebohnen', 'arabica', 'espresso', 'robusta', 'fairtrade', 'bohne', 'kakao'	0.8 %
35	water	koi', 'teich', 'aquarium', 'algen', 'fische', 'wasser', 'pflanzen', 'wasserpflanzen', 'gartenteich', 'futter'	0.7 %
36	cosmetics & food supplements	protein', 'aminosäuren', 'körper', 'eiweiß', 'ernährung', 'proteine', 'soja', 'whey', 'veganer', 'vitamin'	0.7 %
37	cosmetics & food supplements	cbd', 'cannabis', 'thc', 'hanf', 'öl', 'cannabidiol', 'hanfpflanze', 'cannabinoide', 'wirkung', 'blüten'	0.7 %
38	agriculture	rasen', 'rollrasen', 'mähen', 'rasenfläche', 'kunstrasen', 'vertikutieren', 'rasens', 'unkraut', 'moos', 'grün'	0.7 %
39	food	tee', 'tees', 'mate', 'geschmack', 'blätter', 'tasse', 'bio', 'matcha', 'aroma', 'grüntee'	0.7 %
40	food	öl', 'fettsäuren', 'omega', 'samen', 'öle', 'kokosöl', 'leinöl', 'rapsöl', 'gewonnen', 'olivenöl'	0.6 %
41	wood	holzspalter', 'stihl', 'forst', 'motorsäge', 'gerät', 'geräte', 'akku', 'motorsägen', 'arbeiten', 'häcksler'	0.6 %
42	agriculture	kartoffeln', 'kartoffel', 'anbau', 'gemüse', 'zwiebeln', 'knollen', 'pommes', 'sorten', 'speisekartoffeln'	0.5 %
43	food	grill', 'grillen', 'bbq', 'holzkohle', 'smoker', 'grillgut', 'fleisch', 'grills', 'temperatur', 'räuchern'	0.5 %
44	wood	bett', 'holz', 'matratze', 'betten', 'massivholz', 'schlafzimmer', 'lattenrost', 'kopfteil', 'liegefläche', 'höhe'	0.4 %
45	construction	dachbegrünung', 'dach', 'begrünung', 'gründach', 'dachbegrünungen', 'extensive', 'dächer', 'begrünte', 'pflanzen', 'vorteile'	0.4 %
46	waste and transport	paletten', 'europaletten', 'transport', 'ippc', 'holzpaletten', 'palette', 'kisten', 'holz', 'ispm', 'holzverpackungen'	0.4 %
47	waste and transport	altholz', 'entsorgung', 'container', 'entsorgen', 'abfälle', 'verwertung', 'holz', 'grünschnitt'	0.4 %
48	wood	friedwald', 'urne', 'verstorbenen', 'asche', 'baumbestattung', 'baum', 'beigesetzt', 'ruheforst', 'beisetzung', 'urnen'	0.4 %
49	food	pilze', 'pilz', 'vitalpilze', 'pilzen', 'shiitake', 'pulver', 'vitalpilz', 'reishi', 'vitalpilzen', 'champignons'	0.3 %
50	wood	weihnachtsbaum', 'weihnachtsbäume', 'baum', 'nadeln', 'tanne', 'nordmanntanne', 'weihnachten', 'weihnachtsbäumen', 'tannenbaum', 'christbaum'	0.3 %
51	construction	kork', 'korkboden', 'korkböden', 'rinde', 'korkeiche', 'bodenbelag', 'elastisch', 'boden', 'eigenschaften', 'linoleum'	0.3 %
52	wood	treppe', 'treppen', 'holztreppen', 'holztreppe', 'stufen', 'handlauf', 'buche', 'geländer', 'holz', 'eiche'	0.3 %
53	food	bier', 'hopfen', 'malz', 'hefe', 'brauerei', 'biere', 'würze', 'brauen', 'gerste', 'gebraut'	0.3 %
54	wood	klang', 'instrument', 'hölzer', 'instrumente', 'ahorn', 'mahagoni', 'holz', 'ebenholz', 'korpus', 'gefertigt'	0.3 %
55	wood	sauna', 'saunen', 'saunaofen', 'holz', 'mm', 'harvia', 'espe', 'finnischen', 'fichte', 'finnische'	0.2 %

Online Appendix

The online appendix is available at https://osf.io/yvfwh/.

REGION

The Journal of ERSA
Powered by WU

Volume 11, Number 1, 2024, 79–109
DOI: 10.18335/region.v11i1.493

journal homepage: region.ersa.org
ISSN: 2409-5370

Exploring economic activity from outer space: A Python notebook for processing and analyzing satellite nighttime lights

Carlos Mendez[1], **Ayush Patnaik**[2]

[1] Nagoya University, Nagoya, Japan
[2] xKDR Forum, Mumbai, India

Received: 25 April 2023/Accepted: 3 April 2024

Abstract. Nighttime lights (NTL) data are increasingly used as a proxy for monitoring national, subnational, and supranational economic activity. These data offer advantages over traditional economic indicators such as GDP, including greater spatial granularity, timeliness, lower cost, and comparability between regions regardless of statistical capacity or political interference. Despite these benefits, the use of NTL data in regional science has been limited. This is in part due to the lack of accessible methods for processing and analyzing satellite images. To address this issue, this paper presents a user-friendly geocomputational notebook that illustrates how to process and analyze satellite NTL images. The evolution of regional disparities in India is presented as an illustrative example. The notebook first introduces a cloud-based Python environment for visualizing, analyzing, and transforming raster satellite images into tabular data. Next, it presents interactive tools for exploring the space-time patterns of the tabulated data. Finally, it describes methods for evaluating the usefulness of NTL data in terms of their cross-sectional predictions, time-series predictions, and regional inequality dynamics.

Key words: satellite nighttime lights, regional income, zonal statistics, exploratory data analysis, panel data analysis, inequality dynamics, Jupyter notebook

1 Introduction

Nighttime lights (NTL) data have become a widely recognized proxy to monitor economic activity at the national, subnational, and supranational levels (Chen, Nordhaus 2011, Henderson et al. 2012, Sutton et al. 2007). The use of NTL data can offer considerable advantages over traditional economic indicators, such as GDP. For example, NTL data provide greater spatial granularity, are more timely, and are less costly to construct than GDP. Furthermore, NTL data are comparable between multiple regions, regardless of differences in statistical capacity, political interference, or informal activities.

In regional science, several topics could benefit from the use of nighttime light data. One main topic is the study of regional development and inequality. Researchers have used nighttime light data as a proxy for subnational income and have found evidence of increasing or decreasing regional inequality over time (Lessmann, Seidel 2017). These data can also be used to examine the relationship between regional development and various factors, such as ethnic inequality (Alesina et al. 2016). Additionally, nighttime light data can be used to analyze the impact of localized policies (Kim et al. 2024). Overall, nighttime light data provide further opportunities to study economic performance and development at the subnational level, particularly in regions where reliable data are difficult to obtain (see the survey article of Gibson et al. (2020) for a more comprehensive overview the use of NTL in economics and Zheng et al. (2023) for an overview of urban applications).

Despite their potential benefits, the latest NTL data products (Elvidge et al. 2017, 2021, Li et al. 2020, Román et al. 2018) have had limited use in the regional science literature. One plausible reason is the lack of accessible methods for processing and analyzing satellite images. Specifically, the processing of large raster-based satellite images into tabular data has made it difficult for researchers to use the latest satellite data products. To address this issue, we introduce a geocomputational notebook that provides a step-by-step guide on how to process and analyze recent satellite NTL images. To illustrate the functionality of this notebook, we study the evolution of regional disparities in India. Rapid economic growth, coupled with notable regional imbalances in India, can be an interesting topic for evaluating the potential uses and limitations of nighttime light data.

Accessing, processing, and analyzing satellite data requires a basic understanding of remote sensing and programming principles. To lessen the learning curve and encourage the use of these data, a growing number of researchers have been sharing their data processing routines as libraries or functions across several programming languages. For instance, Falchetta (2023), Miethe (2023) and Njuguna (2020) have developed R libraries dedicated to extracting data from nightlight satellite imagery. Raschky (2020) has provided Python functions for the same purpose. Patnaik et al. (2023) have assembled a Julia library that specializes in processing nightlight imagery. With the same motivation, our contribution lies in presenting an alternative, yet complementary, approach for processing and analyzing satellite imagery. By leveraging the pedagogical and computational features of Jupyter notebooks (Rowe et al. 2020, Reades 2020, Chen et al. 2020), our cloud-based geocomputational Jupyter notebook provides an integrated environment for efficient code execution, interactive data visualization, and narrative documentation.

The notebook begins by introducing a cloud-based Python environment for visualizing and transforming raster images into tabular data. The notebook then presents interactive tools to explore the space-time patterns of the tabulated data. These tools allow researchers to better understand both the spatial distribution and the temporal trends of NTL data. To develop a sense of the informational content of NTL, the space-time patterns of GDP are also presented. Finally, the notebook illustrates methods for evaluating the usefulness of NTL data in terms of cross-sectional predictions, time-series predictions, and regional inequality dynamics.

2 Cloud-based environment

Modern computational notebooks allow us to present code in conjunction with descriptive text, equations, visualizations, and tables in a single document (Rowe et al. 2020). The use of such notebooks greatly enhances the reproducibility and transparency of scientific research. Despite the advances offered by computational notebooks, a significant challenge persists in the reproducibility of the computational environment, which is essential for generating consistent results. Especially for geospatial analysis, a notebook user still needs to download, install, and manage numerous computational libraries and their dependencies.

Cloud-based environments such as Google Colab, Anaconda Cloud or Deepnote offer solutions to the reproducible-environment problem. They operate on cloud computers that can be reproduced with a single click. To process and analyze satellite images in a fully reproducible cloud-based environment, we host our notebook on Google Colab: https://colab.research.google.com/github/quarcs-lab/project2022p/blob/master/-project2022p_notebook.ipynb. This cloud-based environment can be easily duplicated, run, and extended after logging in with a Google account. Furthermore, when operating in the cloud, the "forms" feature of Google Colab's code cells facilitates the display, folding,

Table 1: List of required packages and standard libraries

Package	Version	Description
numpy	1.23.5	Library that provides functions for mathematical operations and handling arrays
pandas	1.5.3	Library that provides a data frame class and functions to manipulate data frames
geopandas	0.13.2	Library that helps work with spatial data
matplotlib	3.7.1	Plotting library, including plotting functions
contextily	1.3.0	Library that helps to add base layers to maps
rasterio	1.3.9	Library for raster data processing
linearmodels	4.27	Library for linear regressions, including panel data analysis
inequality	1.0.0	Library that provides methods for measuring inequality
os		Operating system interface
requests	2.31.0	HTTP library for making requests in Python
glob		File path pattern matching
shutil		High-level file operation utilities
bs4		BeautifulSoup library for parsing HTML and XML
json		Library for working with JSON data
gzip		Library for compressing and decompressing files using the gzip format
plotly	5.15.0	High-level library for creating interactive visualizations with Plotly
cufflinks	0.17.3	Productivity Tools for Plotly + Pandas
mpl_toolkits		Tool for advanced axes layout in Matplotlib
linearmodels	4.27	Library for performing linear regressions

and parameterization of code.

At its minimum, this environment requires the libraries shown in Table 1. The results were produced using Python 3.10.12. To start, we install the required Python libraries that are not pre-installed in Google Colab.

[1]:

```
# @title CODE: Install libraries

!pip install \
numpy==1.23.5 \
pandas==1.5.3 \
geopandas==0.13.2 \
matplotlib==3.7.1 \
contextily==1.3.0 \
rasterio==1.3.9 \
folium==0.14.0 \
kaleido==0.2.1 \
mapclassify==2.6.0 \
linearmodels==4.27 \
inequality==1.0.0 \
cufflinks==0.17.3 \
requests==2.31.0 \
plotly==5.15.0 \
--quiet
```

In the next cell, we load the libraries.

[ ]:

```
# @title CODE: Import libraries

import numpy as np        # Library that provides functions for mathematical operations
                          # and handling arrays
import pandas as pd       # Library that provides a data frame class and functions to
                          # manipulate data frames
import geopandas as gpd   # Library that helps working with spatial data


import matplotlib.pyplot as plt  # Function for 2D plotting
from matplotlib.gridspec import GridSpec
from mpl_toolkits.axes_grid1 import (
    make_axes_locatable,
)  # Function to create a new axis on a plot

import contextily as cx   # Library that helps adding OSM base layer to plots.
import plotly.express as px  # Library for interactive plotting

# Libraries to allow plotly to be offline and show plots in a jupyter notebook
import plotly.io as pio
import plotly.graph_objects as go
import cufflinks as cf

cf.go_offline()

import rasterio  # Library for raster data processing
from rasterio import plot as rioplot  # Function to plot raster data
from rasterio.mask import (
    mask,
)  # Function for masking raster data using shapefile for zonal statistics

# linearmodels library provides helps performing regressions
from linearmodels import PooledOLS  # Function to perform pooled OLS regression
from linearmodels import PanelOLS  # Function to perform OLS regression on panel data
from linearmodels import (
    BetweenOLS,
)  # Function to compute the between estimator of an OLS regression
from linearmodels.panel.results import (
    compare,
)  # Function compare results of an OLS regression

import inequality  # Library that provides methods for measuring spatial inequality

import os  # Operating system interface
import requests  # HTTP library for making requests in Python
import glob  # File path pattern matching
import shutil  # High-level file operation utilities
from bs4 import BeautifulSoup  # BeautifulSoup library for parsing HTML and XML
import json  # Library for working with JSON data
import gzip  # Library for compressing and decompressing files using the gzip format
```

We configure table display parameters to abbreviate content and ensure readability. Additionally, we adjust the Plotly renderer to “colab” when the notebook is executed in that environment.

[3]:

```
# @title CODE: Set parameters

# Abbreviate displayed tables

# Set precision to 4
pd.set_option("display.precision", 4)

# Set max columns to 7
pd.set_option("display.max_columns", 7)

# Set max rows to 10
pd.set_option("display.max_rows", 10)

# Set renderer for plotly

if 'COLAB_GPU' in os.environ:
    pio.renderers.default = "colab"
elif 'NOTEBOOK_MODE' in os.environ and os.environ['NOTEBOOK_MODE'] == 'colab':
    pio.renderers.default = "colab"
```

3 Data

In this notebook, we use three datasets: (1) satellite nighttime light images from Elvidge et al. (2021); (2) subnational income per capita from Smits, Permanyer (2019); and (3) administrative boundaries for the states of India from Smits, Permanyer (2019). To work with these datasets, we first need to define our analysis period and organize the directory structure of our computational environment. We define the start and end years as global variables. This definition restricts the study to a particular time frame, allowing us to focus on a particular period of interest. The start and end years can take values from 2014 to 2021.

Note: While minimal adjustments enable the use of images of 2012 and 2013, it's important to acknowledge that they have not undergone stray light correction.

[4]:

```
# @title CODE: Define start and end years
START_YEAR = 2014  # @param {type:"integer"}

END_YEAR = 2019  # @param {type:"integer"}
```

Next, we define the directory paths where the datasets will be downloaded. In the same cell, we also define the paths where figures and tables will be saved during the execution of the notebook. Appendix B gives more details about these folders and the files which will be saved in them.

We also create the directories and delete the `sample_data` directory of Colab.

[5]:

```
# @title CODE: Define paths and directories
# Remove sample_data folder provided by Colab.
!rm -rf sample_data
# Define path constants
FIGURES_DIRECTORY = "figures"
TABLES_DIRECTORY = "tables"
VECTOR_DIRECTORY = "data/vector"
TABULAR_DIRECTORY = "data/tabular"
RASTER_DIRECTORY = "data/raster"

# Check and create folders using a loop
for path in [
    FIGURES_DIRECTORY,
    TABLES_DIRECTORY,
    VECTOR_DIRECTORY,
    TABULAR_DIRECTORY,
    RASTER_DIRECTORY,
]:
    if not os.path.exists(path):
        os.makedirs(path)
        print(f"Created     folder: {path}")
```

3.1 Satellite nighttime light images

The Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) and the Visible Infrared Imaging Radiometer Suite (VIIRS) are two satellite sensors that have been used to construct global nighttime lights datasets (Levin et al. 2020, Elvidge et al. 2013, Gibson et al. 2020, Donaldson, Storeygard 2016). The DMSP-OLS, launched in the 1970s by the US Air Force, was the first sensor used to systematically collect low-light imaging data of the earth at night. The resulting DMSP nighttime lights dataset has been widely used since the 1990s to study socioeconomic activity (Henderson et al. 2012, Chen, Nordhaus 2011, Gibson 2020). In contrast, the VIIRS sensor was launched by NASA and NOAA in 2011 on the Suomi NPP satellite. In remote sensing and natural sciences, the higher-quality VIIRS data quickly began to replace DMSP data in many applications (Elvidge et al. 2013, 2017, 2021). In economics and social sciences, however, the use of DMSP data is still predominant, and concerns have been raised about the measurement errors associated with these data (Abrahams et al. 2018, Gibson et al. 2021, Gibson 2020, Zhang et al. 2023).

The VIIRS nighttime light data has several advantages over the DMSP data. The VIIRS dataset has a much finer spatial resolution, 15 arc-second grids compared to the 30 arc-second grids for DMSP. The measurement units of the VIIRS data are radiance values in radiometric units, specifically nanowatts per square centimeter per steradian (nW/cm2/sr). On the other hand, DMSP data employ radiance values measured in digital numbers (DN). Both higher resolution and more precise measurement units allow the VIIRS data to provide a more accurate delineation between urban centers and rural areas. In addition, the VIIRS sensors have a greater radiometric range, better calibration, and less burring, providing more consistent measurements over time and across space. However, VIIRS data currently have shorter time series than the DMSP data, spanning only 2012-present compared to 1992-2013 for DMSP. This makes the VIIRS data less useful for studying long-term economic and urban trends. Overall, the VIIRS data offer superior features for cross-sectional and time series analyses, but researchers requiring long-time series may still need to use the DMSP data despite their flaws.

In this notebook, we focus on analyzing the annual VIIRS 2.1 dataset, as pre-processed by Elvidge et al. (2021). These authors employed monthly averages of radiance without cloud interference to create a series of global nighttime light images spanning 2012-2019. The pre-processing steps involved the exclusion of biomass burning, aurora, and background noise. Outliers were eliminated through a twelve-month median. Areas lacking detectable lighting were identified using a statistical texture measure. Additionally, the images exhibit enhanced noise filtering, attributed to an extended threshold applied over multiple years. As a result of this pre-processing, the images provide high spatial-temporal consistency and facilitate the analysis of changes over time. The raster images associated with this dataset are available in the website of Earth Observation Group (EOG) of the Payne Institute for Public Policy of the Colorado School of Mines: https: //eogdata.mines.edu/products/vnl. For complementary and exploratory purposes, Appendix C briefly illustrates how to access and analyze the calibrated version of the DMSP dataset that has been pre-processed by Li et al. (2020). The raster images associated with this dataset (Li et al. 2021) are available in Figshare: https://doi.org/10.6084/m9.figshare.9828827.v5.

The VIIRS nighttime lights data can be downloaded via EOG's API by creating a free account at the registration page: https://eogdata.mines.edu/products/register/. After registration, the values of the USERNAME and PASSWORD need to be updated in the following cell. The CLIENT_ID and CLIENT_SECRET are public, but they may change from time to time, and if the cell fails to run, the values should be updated. The latest values can be found on EOG's registration page.

[6]:

```
# @title CODE: Write your EOG's access credentials

## Make an account at EOG: https://eogdata.mines.edu/products/register/
USERNAME = "" # @param {type:"string"}
PASSWORD = ""  # @param {type:"string"}
CLIENT_ID = "eogdata_oidc" # @param {type:"string"}
CLIENT_SECRET = "2677ad81-521b-4869-8480-6d05b9e57d48" # @param {type:"string"}


def download_link(link):
    """
    Function to download image from EOG using the API

    Parameters:
    -----------
    link : str

    Returns:
    --------
    None

    Notes:
    ------
    Downloads image in RASTER_DIRECTORY
    """
    output_file = os.path.basename(link)
    output_file = os.path.join(
        RASTER_DIRECTORY, output_file
    )  # Construct full path with folder
    if os.path.isfile(output_file):
        print(f"File {output_file} already exists. Skipping download.")
        return

    params = {
        "client_id": CLIENT_ID,
        "client_secret": CLIENT_SECRET,
        "username": USERNAME,
        "password": PASSWORD,
        "grant_type": "password",
    }
    token_url = (
        "https://eogauth.mines.edu/auth/realms/master/protocol/openid-connect/token"
    )
    response = requests.post(token_url, data=params)
    access_token_dict = json.loads(response.text)
    access_token = access_token_dict.get("access_token")
    data_url = link
    auth = "Bearer " + access_token
    headers = {"Authorization": auth}
    response = requests.get(data_url, headers=headers)

    with open(output_file, "wb") as f:
        f.write(response.content)
    print(f"File {output_file} downloaded successfully.")


base_url = "https://eogdata.mines.edu/nighttime_light/annual/v21/{}/"

for year in range(
    START_YEAR, END_YEAR + 1
):  # loop through years START_YEAR to END_YEAR
    url = f"https://eogdata.mines.edu/nighttime_light/annual/v21/{year}/VNL_v21_npp_
            {year}_global_vcmslcfg_c202205302300.median_masked.dat.tif.gz"
    download_link(url)
```

[6]:

```
File data/raster\VNL_v21_npp_2014_global_(...).dat.tif.gz downloaded successfully.
File data/raster\VNL_v21_npp_2015_global_(...).dat.tif.gz downloaded successfully.
File data/raster\VNL_v21_npp_2016_global_(...).dat.tif.gz downloaded successfully.
File data/raster\VNL_v21_npp_2017_global_(...).dat.tif.gz downloaded successfully.
File data/raster\VNL_v21_npp_2018_global_(...).dat.tif.gz downloaded successfully.
File data/raster\VNL_v21_npp_2019_global_(...).dat.tif.gz downloaded successfully.
```

The loop runs from START_YEAR to END_YEAR and fetches the files containing VIIRS V2.1 images for each year. The downloaded files are compressed, and need to be extracted to be used. For illustration and computational efficiency purposes, we focus only on the 2014-2019 period. An analysis for other years (2012, 2020-onwards) is also possible by adjusting the names of the decompressed files in a sequential way.

[7]:

```
# @title CODE: Extract downloaded images

# Get the current working directory
current_dir = os.getcwd()

# Find all .gz files in the data/raster folder

for file in glob.glob(os.path.join(current_dir, RASTER_DIRECTORY, "*.gz")):
    with gzip.open(file, "rb") as compressed_file, open(
        file[:-3], "wb"
    ) as extracted_file:
        shutil.copyfileobj(compressed_file, extracted_file)  # Stream data directly
    os.remove(file)

print("Successfully extracted all .gz files!")
```

3.2 Regional income and administrative boundaries

Smits, Permanyer (2019) have recently compiled a database of socioeconomic indicators at the subnational level. The dataset is based on the first-level administrative regions. As an indicator of subnational GDP per capita, we use Gross National Income per capita in thousands of US dollars (2011 PPP) from this database. We utilize version 4.0 of the dataset, which also includes a shapefile containing the boundaries of the administrative regions. All these datasets are available from the Global Data Lab website: https://globaldatalab.org.

A small part of the dataset, for the Indian states, has been downloaded and uploaded to GitHub to facilitate the exposition of this notebook. In the code below, the subnational GDP file of India is loaded as a pandas dataframe.

[8]:

```
# @title CODE: Read state level GDP data from a csv file

df_GDP = pd.read_csv(
    "https://gist.github.com/cmg777/150c0b93ae8eb14fec9babdf4f5f8fc4/raw/
             2af006ed2b80cdb3ef9b6dd13ccc8a450c168765/df_GDP_India36.csv"
)
```

Next, we load the vector file that contains the borders of Indian states using `geopandas`. The file is loaded into a `GeoDataFrame`, which has a column that contains geometric information in the form of polygons. This column is utilized for conducting geometric operations, such as spatial joins and intersections, as well as for visualizing the data using tools like `matplotlib`.

[9]:

```
# @title CODE: Read administrative boundaries

# The boundaries of states of India are stored in gdf_india36.geojson. The file is
loaded using the read_file function from geopandas
map_url = "https://gist.github.com/cmg777/19c25af8fcfe2291cfb6f9abf141d45a/raw/
                   48e1489e97f975c5a2253d2068cf99a3c2d0cff3/gdf_india36.geojson"

polygons_files = gpd.read_file(map_url)
```

4 Processing satellite nighttime images

4.1 Importing and visualizing satellite images

We define a function to load a satellite image (raster file) for a given year. The function finds the path of the file corresponding to the provided year value, and loads it using `rasterio`.

[10]:

```
# @title CODE: Define function to load satellite images


def load_raster(year):
    """
    Load a raster file based on the provided time identifier.

    Parameters:
    -----------
    year : int

    Returns:
    --------
    rasterio.io.DatasetReader
    An opened raster file dataset ready for further operations.

    Example:
    --------
    >>> raster_2014 = load_raster(2014)
    >>> type(raster_2014)
    <class 'rasterio.io.DatasetReader'>

    Notes:
    ------
    Modify the path in the function if your file structure
    or naming convention differs.
    """
    raster_path = f"{RASTER_DIRECTORY}/VNL_v21_npp_{year}*"
    return rasterio.open(glob.glob(raster_path)[0])
```

We next extract the bounding box of the country map (vector file), which is the latitude and longitude extent of India. This allows us to crop the NTL images to values only for our area of interest. This step can reduce memory usage, computation time and make it easier to visualise the NTL around the area of interest.

[11]:

```
# @title CODE: Extract the bounding box using the boundaries of the map

polygons_files_bbox = polygons_files.total_bounds
```

Next, we utilize `load_raster` to load the NTL images of the world for the start and end years. Then, we employ the bounding box of the vector file to crop the image specifically to India.

[12]:

```
# @title CODE: Load images and crop them

# The raster file corresponding to the start year is loaded using the open function
# in rasterio
raster_file_first = load_raster(START_YEAR)

# The bounding box of the vector data is used to crop the raster file
raster_file_window_first = raster_file_first.window(*polygons_files_bbox)
raster_file_clipped_first = raster_file_first.read(1, window=raster_file_window_first)

# The raster file corresponding to the end year is loaded using the open function
# in rasterio
raster_file_last = load_raster(END_YEAR)

# The bounding box of the vector data is used to crop the raster file
raster_file_window_last = raster_file_last.window(*polygons_files_bbox)
raster_file_clipped_last = raster_file_last.read(1, window=raster_file_window_last)
```

In Figure 1, we overlay the cropped NTL images for the start and end years. We then superimpose the state boundaries from the vector file. As expected, India exhibits a brighter appearance in the image for the end year in comparison to the start year.

[13]:

```
# @title CODE: Plot satellite images before and after

# Initialize the GridSpec for setting up the plot structure
RADIANCE_THRESHOLD = 6  # @param {type:"number"}

gs = GridSpec(1, 3, width_ratios=[2, 2, 0.1])

fig = plt.figure(figsize=(15, 8))
ax1 = fig.add_subplot(gs[0])

first_year_plot = ax1.imshow(
    raster_file_clipped_first,
    extent=polygons_files_bbox[[0, 2, 1, 3]],
    vmin=0,
    vmax=RADIANCE_THRESHOLD,
    cmap="magma",
)
polygons_files.boundary.plot(ax=ax1, color="skyblue", linewidth=0.4)
cx.add_basemap(ax1, crs=polygons_files.crs.to_string(),
               source = cx.providers.CartoDB.DarkMatterOnlyLabels, attribution=False)

ax1.set_title(f"(a) Nighttime lights in {START_YEAR}")
ax1.set_axis_off()

ax2 = fig.add_subplot(gs[1])

last_year_plot = ax2.imshow(
    raster_file_clipped_last,
    extent=polygons_files_bbox[[0, 2, 1, 3]],
    vmin=0,
    vmax=RADIANCE_THRESHOLD,
    cmap="magma",
)
polygons_files.boundary.plot(ax=ax2, color="skyblue", linewidth=0.4)
cx.add_basemap(ax2, crs=polygons_files.crs.to_string(),
               source = cx.providers.CartoDB.DarkMatterOnlyLabels, attribution=False)
ax2.set_title(f"(b) Nighttime lights in {END_YEAR}")
ax2.set_axis_off()

cax2 = fig.add_subplot(gs[2])
# Add colorbar
cbar = fig.colorbar(
    first_year_plot, cax=cax2, label="Luminosity intensity (nanoWatts/sr/$cm^2$)"
)
```

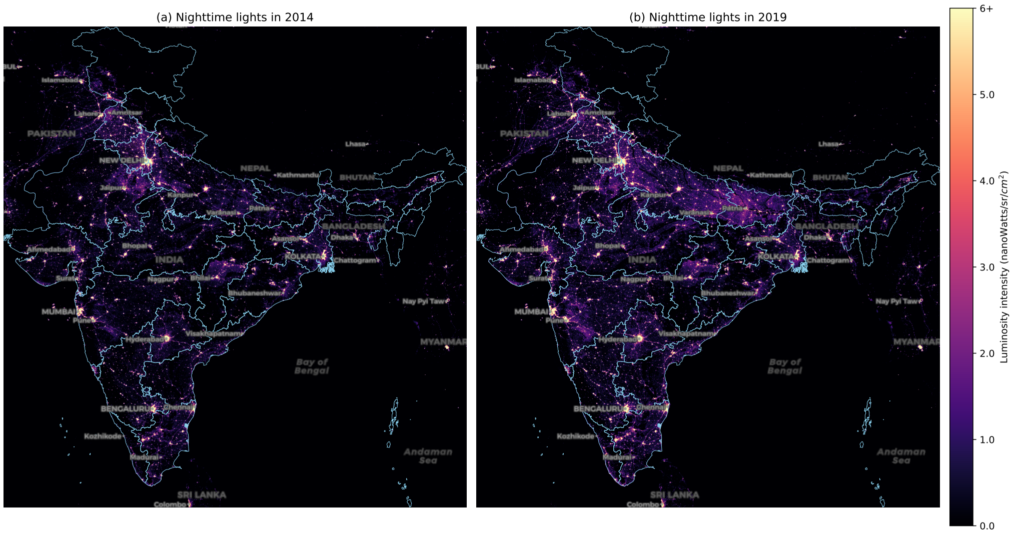


Figure 1: Raster images of nighttime lights and administrative boundaries of India: Initial vs final year

```
# Set ticks and labels with the last one as f"{RADIANCE_THRESHOLD}+"
cbar.set_ticklabels(
    [f"{val}" for val in cbar.get_ticks()[:-1]] + [f"{RADIANCE_THRESHOLD}+"]
)

plt.tight_layout()
plt.savefig("figures/NTL.png", dpi=300, bbox_inches="tight")
plt.show()
```

[13]:
```
Output in Figure 1
```

4.2 Computing zonal statistics

To make a meaningful comparison between NTL and subnational GDP, it is crucial to ensure that both datasets are measured at the same geographic scale. As GDP is available at the state level, we need to aggregate the NTL values accordingly. This can be achieved through zonal statistics, where we sum the amount of light within the boundaries of each state, resulting in a state-level dataset of regional luminosity.

To accomplish this aggregation, we initially load the NTL images for each year of interest. Next, we define a mask function that can filter all points outside the polygon in the raster file. Lastly, we apply the mask function to each polygon in the vector file, resulting in the summation that generates state-level luminosity for each year. More details about this implementation are provided in Appendix A. Although missing values are not a problem in our state-level dataset, they are more prone to occur when geographical units are small, largely rural, or sparsely populated.

[14]:
```
# @title CODE: Define dataset to store results

gdf_NTL = polygons_files.copy()
```

[15]:
```
# @title CODE: Compute zonal statistics

# Choose an operator for aggregation. In this notebook, the operator, AGGREGATE_OPERATOR,
# has been set to np.ma.sum.
# Other operators can be chosen, for example, np.ma.mean and np.ma.median will compute
# the mean and median respectively.
# The list of operators can be found here:
#     https://numpy.org/doc/stable/reference/routines.ma.html
AGGREGATE_OPERATOR = np.ma.sum
```

Table 2: Regional nighttime light values over time

	id	region	geometry	...	2017	2018	2019
0	1	Andaman and Nicobar Islands	MULTIPOLYGON (((93.84...	...	2.9501e+03	3.3096e+03	3.4267e+03
1	3	Arunachal Pradesh	MULTIPOLYGON (((95.23...	...	8.4099e+03	8.1394e+03	1.1512e+04
2	4	Assam	MULTIPOLYGON (((95.19...	...	1.7227e+05	1.6381e+05	1.6768e+05
3	5	Bihar	MULTIPOLYGON (((88.11...	...	4.4029e+05	5.3907e+05	6.2007e+05
4	6	Chandigarth	MULTIPOLYGON (((76.84...	...	1.3767e+04	1.3751e+04	1.4122e+04
...	...	...	...	...	...	...	...
31	34	Uttar Pradesh	MULTIPOLYGON (((79.39...	...	1.7540e+06	1.6793e+06	1.6754e+06
32	35	Uttarakhand	MULTIPOLYGON (((80.07...	...	1.1927e+05	1.1099e+05	1.1275e+05
33	36	West Bengal	MULTIPOLYGON (((88.49...	...	5.2243e+05	5.1658e+05	5.0112e+05
34	26	Odisha	MULTIPOLYGON (((86.72...	...	3.6816e+05	3.7614e+05	3.6962e+05
35	2	Andhra Pradesh	MULTIPOLYGON (((81.10...	...	3.7570e+05	4.3301e+05	4.6353e+05

```
# Define the clean_mask function outside the loop
def geom_mask(geom, dataset, crop=True, all_touched=True):
    masked, mask_transform = mask(
        dataset=dataset, shapes=(geom,), crop=crop, all_touched=all_touched
    )
    return masked


# A loop runs from the start year to the end year that computes the aggregate nighttime
# lights radiance for each state
for year in range(START_YEAR, END_YEAR + 1):
    # The raster file of the given year is loaded
    raster_file = load_raster(year)
    # The mask is applied, and then a summation is performed for computing the aggregate
    # radiance.
    statewise_agg_ntl = polygons_files.geometry.apply(
        geom_mask, dataset=raster_file
    ).apply(AGGREGATE_OPERATOR)

    # The state-wise aggregate radiance of the year is stored in the data frame that was
    # initialized earlier.
    gdf_NTL[str(year)] = statewise_agg_ntl
```

As the final step in the aggregation process, we obtain a `GeoDataFrame` in which each column represents the total sum of nighttime lights for each state across all years.

[16]:

```
# @title CODE: Show zonal statistics results
gdf_NTL
```

[16]:

```
                Output in Table 2
```

Next, we create a dataset of summary statistics for state-level nighttime lights by year.

[17]:

```
# @title CODE: Show descriptive statistics by year

# The summary statistics of aggregate nighttime lights is produced for each year
gdf_NTL_summary = gdf_NTL.drop(["geometry", "id"], axis=1).describe().round(2)
gdf_NTL_summary
```

[17]:

```
                Output in Table 3
```

Next, we export the results as a geojson file, ready for analysis in standard geospatial applications and data science languages.

[18]:

```
# @title CODE: Save tabular NTL dataset

gdf_NTL.to_file("data/vector/gdf_NTL.geojson", driver="GeoJSON")
```

4.3 Creating panel-data structures

Having successfully aggregated the nighttime lights data to the state level to align with the resolution of the subnational GDP data, we can now create panel-data structures for

Table 3: Descriptive statistics of regional nighttime lights

	2014	2015	2016	2017	2018	2019
count	3.6000e+01	36.00	3.6000e+01	3.6000e+01	3.6000e+01	3.6000e+01
mean	2.4044e+05	254096.11	2.5798e+05	3.0650e+05	3.1984e+05	3.3252e+05
std	2.7458e+05	284386.91	3.0769e+05	3.8398e+05	3.9127e+05	4.0660e+05
min	8.7400e+01	92.03	7.3820e+01	1.0994e+02	1.2033e+02	1.3022e+02
25%	1.1262e+04	11442.40	1.2232e+04	1.3355e+04	1.3229e+04	1.3640e+04
50%	1.4720e+05	170297.48	1.4774e+05	1.6523e+05	1.6215e+05	1.6067e+05
75%	4.1105e+05	378816.13	3.6443e+05	4.4962e+05	5.1787e+05	5.0807e+05
max	1.0247e+06	993814.31	1.2645e+06	1.7540e+06	1.6793e+06	1.6754e+06

both datasets and merge them into a single dataset. We first retrieve the state-level NTL and GDP data. Both datasets are transformed into long-form panel structures. Thus, they are ready to be merged into a single dataset.

[19]:
```
# @title CODE: Read saved NTL data

df_NTL = gpd.read_file("data/vector/gdf_NTL.geojson").drop("geometry", axis=1)
```

[20]:
```
# @title CODE: Reshape NTL data into long-form panel data

df2_NTL = pd.melt(
    df_NTL,
    id_vars=["id", "region"],
    value_vars=[str(x) for x in range(START_YEAR, END_YEAR + 1)],
)
df2_NTL.columns = ["id", "region", "year", "NTL"]
```

[21]:
```
# @title CODE: Reshape GDP data into long-form panel data

df2_GDP = pd.melt(
    df_GDP,
    id_vars=["id", "region"],
    value_vars=[str(x) for x in range(START_YEAR, END_YEAR + 1)],
)
df2_GDP.columns = ["id", "region", "year", "GDP"]
```

[22]:
```
# @title CODE: Merge NTL and GDP datasets

df = pd.merge(df2_GDP, df2_NTL, on=["id", "region", "year"], how="inner")
```

Two new columns are added on the basis of the natural logarithmic values of NTL and GDP. To avoid calculation problems with the logarithmic values, we add a constant of 0.01.

[23]:
```
# @title CODE: Add offset to compute log values

LOG_OFFSET = 0.01
df["lnNTL"] = np.log(LOG_OFFSET + df["NTL"])
df["lnGDP"] = np.log(LOG_OFFSET + df["GDP"])
```

Finally, we have a long-form panel-data structure of state-level NTL and GDP.

[24]:
```
# @title CODE: Show long-form dataset

df
```

[24]:
```
                Output in Table 4
```

[25]:
```
# @title CODE: Save long-form dataset

df.to_csv("data/tabular/df_NTL_GDP_lnNTL_lnGDP.csv", index=False)
```

For other kinds of analysis, this panel-data structure is reshaped into its wide form. This new dataset contains the following columns: id, region name, geometry, and logarithm values of NTL for each year

Table 4: Regional GDP and nighttime lights: Long-form panel dataset

	id	region	year	GDP	NTL	lnNTL	lnGDP
0	1	Andaman and Nicobar Islands	2014	3.9615e+06	2.9365e+03	7.9850	15.1921
1	2	Andhra Pradesh	2014	4.3554e+08	3.0310e+05	12.6218	19.8921
2	3	Arunachal Pradesh	2014	8.5731e+06	7.2797e+03	8.8928	15.9641
3	4	Assam	2014	1.1240e+08	1.5431e+05	11.9467	18.5376
4	5	Bihar	2014	2.6511e+08	2.0255e+05	12.2187	19.3957
...	...	...	...	...	...	...	...
211	32	Telangana	2019	3.6413e+08	5.2893e+05	13.1786	19.7130
212	33	Tripura	2019	1.9736e+07	3.4391e+04	10.4456	16.7980
213	34	Uttar Pradesh	2019	9.9160e+08	1.6754e+06	14.3316	20.7148
214	35	Uttarakhand	2019	9.9485e+07	1.1275e+05	11.6330	18.4155
215	36	West Bengal	2019	5.5466e+08	5.0112e+05	13.1246	20.1339

[26]:

```
# @title CODE: Generate a wide-form dataset for (ln) NTL

# Pivot panel data from long form to wide form
df_lnNTL = df.pivot_table(
    index=["id", "region"], columns="year", values="lnNTL"
).reset_index(drop=False)
# Make sure the column names are strings
df_lnNTL.columns = df_lnNTL.columns.astype(str)
```

[27]:

```
# @title CODE: Merge (ln) NTL dataset with map file

gdf_lnNTL = pd.merge(
    polygons_files,
    df_lnNTL,
    left_on=["id", "region"],
    right_on=["id", "region"],
    how="inner",
)
gdf_lnNTL
```

[27]:

```
                Output in Table 5
```

Table 5: Regional (ln) NTL: Wide-form panel dataset

	id	region	geometry	...	2017	2018	2019
0	1	Andaman and Nicobar Islands	MULTIPOLYGON (((93.84...	...	7.9896	8.1046	8.1394
1	3	Arunachal Pradesh	MULTIPOLYGON (((95.23...	...	9.0372	9.0045	9.3512
2	4	Assam	MULTIPOLYGON (((95.19...	...	12.0568	12.0064	12.0298
3	5	Bihar	MULTIPOLYGON (((88.11...	...	12.9952	13.1976	13.3376
4	6	Chandigarth	MULTIPOLYGON (((76.84...	...	9.5301	9.5289	9.5555
...	...	...	...	...	...	...	...
31	34	Uttar Pradesh	MULTIPOLYGON (((79.39...	...	14.3774	14.3339	14.3316
32	35	Uttarakhand	MULTIPOLYGON (((80.07...	...	11.6891	11.6172	11.6330
33	36	West Bengal	MULTIPOLYGON (((88.49...	...	13.1662	13.1550	13.1246
34	26	Odisha	MULTIPOLYGON (((86.72...	...	12.8163	12.8377	12.8202
35	2	Andhra Pradesh	MULTIPOLYGON (((81.10...	...	12.8366	12.9785	13.0466

The resulting geospatial dataset is saved and will be used in various visualizations in the next section.

[28]:

```
# @title CODE: Save geospatial dataset of (ln) NTL

gdf_lnNTL.to_file("data/vector/gdf_lnNTL.geojson", driver="GeoJSON")
```

Similarly, we construct a wide-form panel dataset for the natural logarithmic values of GDP.

[29]:

```
# @title CODE: Generate wide-form dataset for (ln) GDP

# Pivot panel data from long form to wide form
df_lnGDP = df.pivot_table(
    index=["id", "region"], columns="year", values="lnGDP"
).reset_index(drop=False)
# Make sure the column names are strings
df_lnGDP.columns = df_lnGDP.columns.astype(str)
```

[30]:

```
# @title CODE: Merge (ln) GDP dataset with map file

gdf_lnGDP = pd.merge(
    polygons_files,
    df_lnGDP,
    left_on=["id", "region"],
    right_on=["id", "region"],
    how="inner",
)
gdf_lnGDP
```

[30]:

```
                Output in Table 6
```

Table 6: Regional (ln) GDP: Wide-form panel dataset

	id	region	geometry	...	2017	2018	2019
0	1	Andaman and Nicobar Islands	MULTIPOLYGON (((93.84...	...	15.2835	15.3464	15.3980
1	3	Arunachal Pradesh	MULTIPOLYGON (((95.23...	...	15.8766	15.9368	15.9862
2	4	Assam	MULTIPOLYGON (((95.19...	...	18.6488	18.7058	18.7527
3	5	Bihar	MULTIPOLYGON (((88.11...	...	19.6709	19.7264	19.7721
4	6	Chandigarth	MULTIPOLYGON (((76.84...	...	16.6551	16.7201	16.7734
...	...	...	...	...	...	...	...
31	34	Uttar Pradesh	MULTIPOLYGON (((79.39...	...	20.6101	20.6675	20.7148
32	35	Uttarakhand	MULTIPOLYGON (((80.07...	...	18.3042	18.3653	18.4155
33	36	West Bengal	MULTIPOLYGON (((88.49...	...	20.0276	20.0859	20.1339
34	26	Odisha	MULTIPOLYGON (((86.72...	...	18.9935	19.0504	19.0973
35	2	Andhra Pradesh	MULTIPOLYGON (((81.10...	...	19.8205	19.8810	19.9307

[31]:

```
# @title CODE: Save geospatial dataset of (ln) NTL

gdf_lnGDP.to_file("data/vector/gdf_lnGDP.geojson", driver="GeoJSON")
```

5 Analyzing nighttime lights and GDP

5.1 Exploring space-time patterns

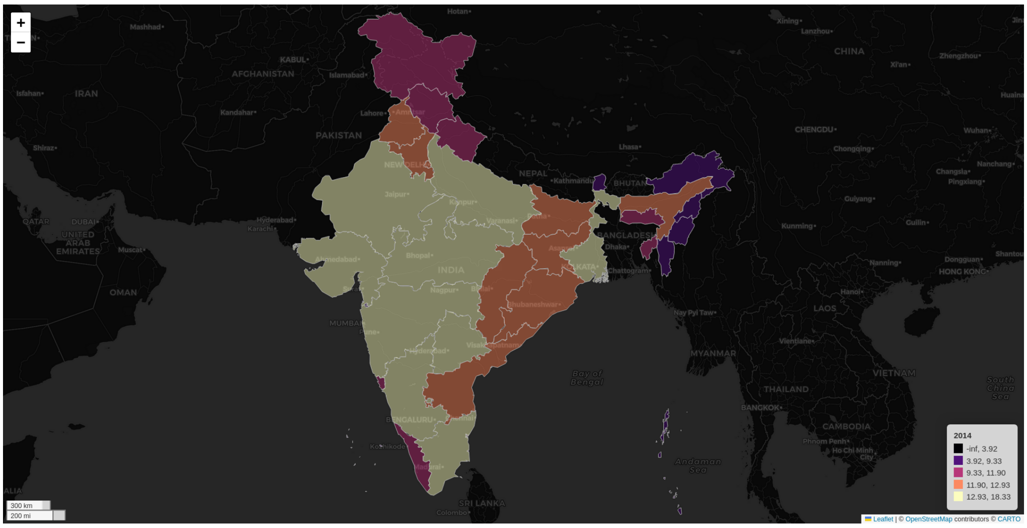
5.1.1 Choropleth maps

Based on the previously constructed panel-data structures (`gdf_lnNTL` and `gdf_lnGDP`), we can easily visualize comparative choropleth maps for (log) nighttime lights and GDP. In particular, the `explore()` function of the Geopandas package allows us to easily construct interactive maps. In addition, consistent with the `Mapclassify` package, multiple classification schemes are available. For example, in Figures 2 and 3, we use a boxplot classification to understand the spatial distribution of NTL and GDP in 2014. We can easily identify where the regions below and above the median are located. We can also identify and compare potential spatial clusters in both distributions.

[32]:

```
# @title CODE: Plot interactive choroplet map for ln NTL

gdf_lnNTL.explore(
    column=str(START_YEAR),
    tooltip=["region", str(START_YEAR)],
    scheme="BoxPlot",  # Quantiles, EqualInterval, BoxPlot, FisherJenks
    cmap="magma",  # hot, cividis, plasma, magma, inferno, coolwarm, viridis
    legend=True,
```



Note: Find the interactive version of this graph at Colab

Figure 2: Distribution of (log) NTL in 2014

```
    tiles="CartoDB dark_matter",  # CartoDB dark_matter OpenStreetMap, Stamen Terrain,
           Stamen Toner, Stamen Watercolor, CartoDB positron, CartoDB dark_mat,
    style_kwds=dict(color="darkgrey", weight=0.8),
    legend_kwds=dict(colorbar=False),
)
```

[32]:
```
                Output in Figure 2
```

[33]:
```
# @title CODE: Plot interactive choroplet map for ln GDP

gdf_lnGDP.explore(
    column=str(END_YEAR),
    tooltip=["region", str(END_YEAR)],
    scheme="BoxPlot",  # Quantiles, EqualInterval, BoxPlot, FisherJenks
    cmap="magma",  # hot, cividis, plasma, magma, inferno, coolwarm, viridis
    legend=True,
    tiles="CartoDB dark_matter",  # CartoDB dark_matter OpenStreetMap, Stamen Terrain,
           Stamen Toner, Stamen Watercolor, CartoDB positron, CartoDB dark_mat,
    style_kwds=dict(color="darkgrey", weight=0.8),
    legend_kwds=dict(colorbar=False),
)
```

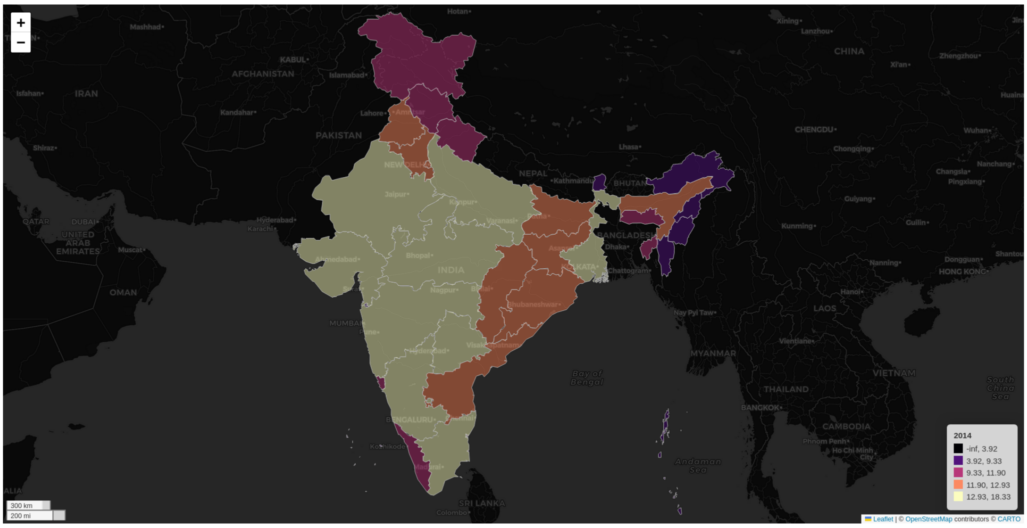
[33]:
```
                Output in Figure 3
```

Static choropleth maps can also be produced, allowing them to be included in non-HTML reports. In Figures 4 and 5, we show how the spatial distributions of NTL and GDP have changed over time. For that purpose, we keep the classification of the initial year constant (except for the minimum and maximum values). Based on these maps, we can observe inter-quantile mobility over time.

[34]:
```
# @title CODE: Plot static map of (log) NTL for the initial and final year

# A figure is initialized
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(12, 8))

# The plot of the start year is added
gdf_lnNTL.plot(
    column=str(START_YEAR),
    scheme="BoxPlot",
    cmap="magma",
    edgecolor="darkgrey",
    legend=True,
    ax=axes[0],
    legend_kwds={"bbox_to_anchor": (0.88, 0.30)},
```



Note: Find the interactive version of this graph at Colab

Figure 3: Distribution of (log) GDP in 2014

```
)
cx.add_basemap(
    ax=axes[0],
    crs=gdf_lnNTL.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterNoLabels,
    attribution=False,
)

cx.add_basemap(
    ax=axes[0],
    crs=gdf_lnNTL.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterOnlyLabels,
    attribution=False,
)

# The plot of the end year is added.
gdf_lnNTL.plot(
    column=str(END_YEAR),
    scheme="user_defined",
    classification_kwds={"bins": [3.92, 9.33, 11.90, 12.93]},
    cmap="magma",
    edgecolor="darkgrey",
    legend=True,
    ax=axes[1],
    legend_kwds={"bbox_to_anchor": (0.88, 0.30)},
)
cx.add_basemap(
    ax=axes[1],
    crs=gdf_lnNTL.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterNoLabels,
    attribution=False,
)

cx.add_basemap(
    ax=axes[1],
    crs=gdf_lnNTL.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterOnlyLabels,
    attribution=False,
)

plt.tight_layout()
axes[0].axis("off")
axes[1].axis("off")
axes[0].set_title("(a) Log of NTL in " + str(START_YEAR))
axes[1].set_title("(b) Log of NTL in " + str(END_YEAR))
```

```
plt.savefig("figures/fig_map_lnNTL.png", dpi=300, bbox_inches="tight")
plt.show()
```

[34]:

```
                    Output in Figure 4
```

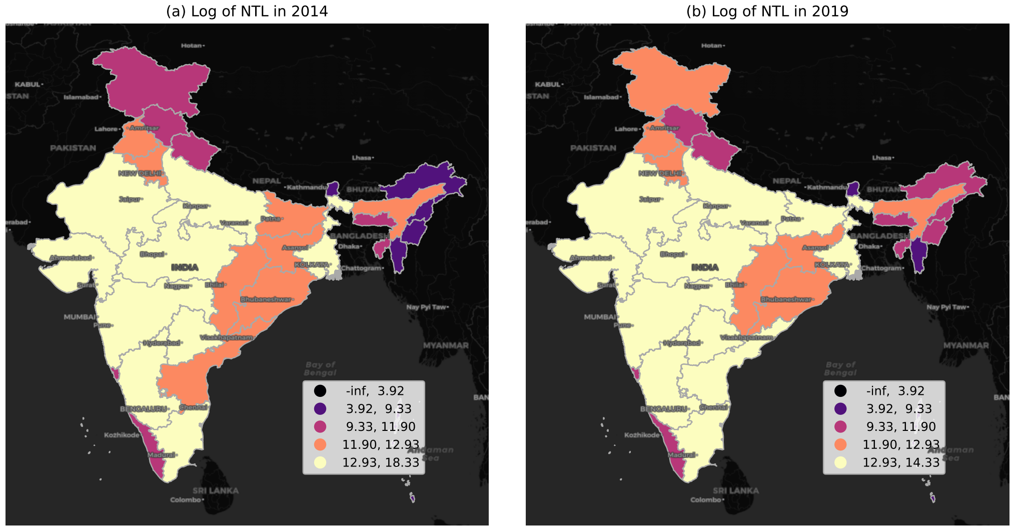


Figure 4: Distribution of (log) nighttime lights: 2014 vs 2019

[35]:

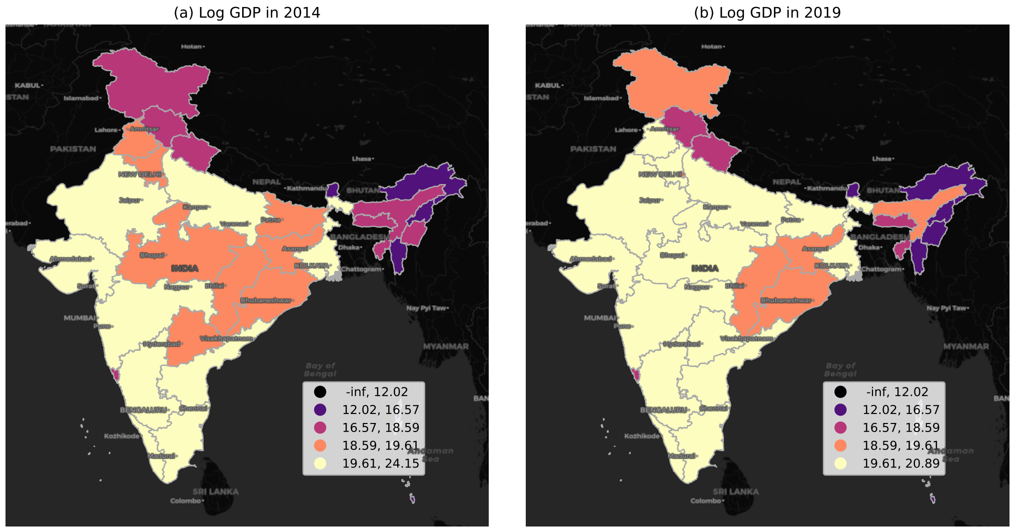
```
#@title CODE: Plot static map of (log) GDP for the initial and final year

fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(12, 8))

# plot for the start year is added.
gdf_lnGDP.plot(
    column=str(START_YEAR),
    scheme="BoxPlot",
    cmap="magma",
    edgecolor="darkgrey",
    legend=True,
    ax=axes[0],
    legend_kwds={"bbox_to_anchor": (0.88, 0.30)},
)
cx.add_basemap(
    ax=axes[0],
    crs=gdf_lnGDP.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterNoLabels,
    attribution=False,
)

cx.add_basemap(
    ax=axes[0],
    crs=gdf_lnGDP.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterOnlyLabels,
    attribution=False,
)

# plot for the end year is added.
gdf_lnGDP.plot(
    column=str(END_YEAR),
    scheme="user_defined",
    classification_kwds={"bins": [12.02, 16.57, 18.59, 19.61]},
    cmap="magma",
    edgecolor="darkgrey",
    legend=True,
    ax=axes[1],
    legend_kwds={"bbox_to_anchor": (0.88, 0.30)},
)
cx.add_basemap(
```


(a) Log GDP in 2014

(b) Log GDP in 2019

Figure 5: Distribution of (log) GDP: 2014 vs 2019

```
    ax=axes[1],
    crs=gdf_lnGDP.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterNoLabels,
    attribution=False,
)

cx.add_basemap(
    ax=axes[1],
    crs=gdf_lnGDP.crs.to_string(),
    source=cx.providers.CartoDB.DarkMatterOnlyLabels,
    attribution=False,
)

plt.tight_layout()
axes[0].axis("off")
axes[1].axis("off")
axes[0].set_title("(a) Log GDP in " + str(START_YEAR))
axes[1].set_title("(b) Log GDP in " + str(END_YEAR))
plt.savefig("figures/fig_map_lnGDP.png", dpi=300, bbox_inches="tight")
plt.show()
```

[35]:

```
                Output in Figure 5
```

5.1.2 Regional time series

In this section, we study the temporal evolution of NTL for each region. As we also have the time series of GDP, we can compare their trends and have a first visual validation of the usefulness of NTL for predicting economic activity over time. The plotting library `Plotly Express` is particularly useful for interactively exploring time series when the dataset is organized as a long-form dataframe. In the code below, we use the previously constructed long-form dataframe (`df`), which contains both NTL and GDP data. After indicating the `x` and `y` variables, we only need to use the argument color to identify the regions. After generating the Plotly object, we use the `write_image()` method to save the results as a static image. To generate a similar graph for GDP, we only need to change one argument: `y = "lnGDP"`.

[36]:

```
# @title CODE: Plot regional time series of (ln) NTL

fig_ts_lnNTL = px.line(df, x="year", y="lnNTL", color="region")
fig_ts_lnNTL.show()
```

[36]:

```
                Output in Figure 6
```

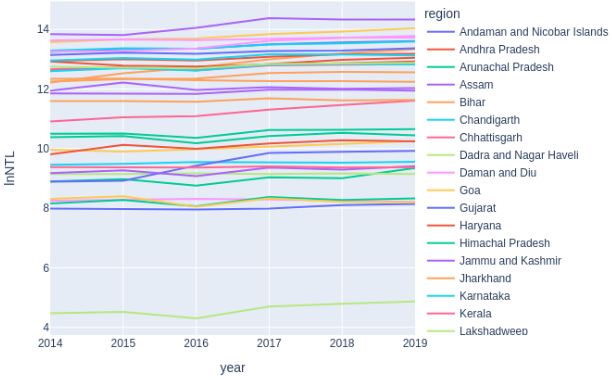


Figure 6: Evolution of (log) nighttime lights in each region

[37]:

```
# @title CODE: Save plotly figure as PNG file

fig_ts_lnNTL.write_image("figures/fig_ts_lnNTL.png")
```

[38]:

```
# @title CODE: Plot regional time series of (ln) GDP

fig_ts_lnGDP = px.line(df, x="year", y="lnGDP", color="region")
fig_ts_lnGDP.show()
```

[38]:

```
                    Output in Figure 7
```

[39]:

```
# @title CODE: Save plotly figure as PNG file

fig_ts_lnGDP.write_image("figures/fig_ts_lnGDP.png")
```

Figures 6 and 7 show the time series of NTL and GDP, respectively, on a regional basis. A preliminary visual examination reveals the similarities and differences between these two variables for each region. In certain regions, the NTL trends exhibit larger fluctuations than those of GDP. Due to the possibility of measurement errors in earth observation data, large fluctuations may require further attention and data processing. For example, to focus the analysis on long-run trends, one may consider using time-series filters to remove short-term fluctuations. Overall, this preliminary visual assessment can provide useful information on the temporal dynamics of economic activity and can be performed effortlessly using the Plotly Express library.

5.1.3 Scatter plot with linear fit

To study the relationship between nighttime lights (NTL) and GDP, we use the interactive scatterplot from the `Plotly Express` library. In addition to the basic arguments of a

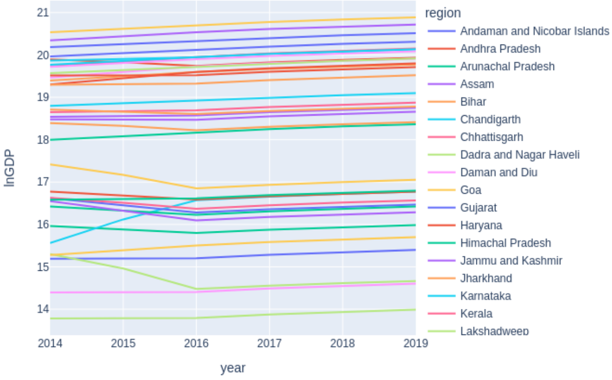


Figure 7: Evolution of (log) GDP in each region

scatter plot (data frame, the x-axis, and the y-axis), the function `px.scatter()` allows us to specify other arguments such as hover on name, text on selected observations, animation frame, and trend line. The animation frame and the trend line options are particularly informative when analyzing longitudinal (panel) data. When activated, we can fit a regression line for each time period. Also, when hovering on the regression line, we can easily obtain regression statistics such as R-Squared, regression coefficients, and predicted values.

[40]:

```
# @title CODE: Plot the NTL-GDP relationship

df_selected_year = df[df["year"].astype(int) == START_YEAR]

N_LABELLED_REGIONS = 6

quantiles = np.linspace(0, 1, N_LABELLED_REGIONS + 1)
quantiles = df_selected_year["lnNTL"].quantile(quantiles)


# Function to find the closest value to a given quantile
def find_closest_value(quantile_val):
    return df_selected_year.iloc[
        (df_selected_year["lnNTL"] - quantile_val).abs().argsort()[:1]
    ]


# Find closest regions to the quantiles
selected_regions = pd.concat([find_closest_value(quantile) for quantile in quantiles])

region_text = selected_regions["region"]

# Create a new column 'selected_region' in df based on the condition
df["selected_regions"] = df["region"].where(df["region"].isin(region_text), pd.NA)

fig_sc_lnNTL_lnGDP = px.scatter(
    data_frame=df,
    x="lnNTL",
    y="lnGDP",
    range_x=[2, 16.5],
    range_y=[12, 22],
    hover_name="region",
    text="selected_regions",
    animation_frame="year",
    trendline="ols",
)

fig_sc_lnNTL_lnGDP.update_traces(textposition="top center")
fig_sc_lnNTL_lnGDP.show()
```

[40]:

```
                Output in Figure 8
```

[41]:

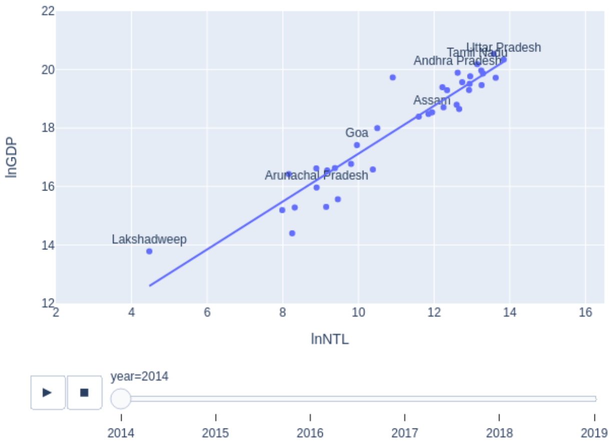
```
# @title CODE: Save plotly figure as PNG

fig_sc_lnNTL_lnGDP.write_image("figures/fig_sc_lnNTL_lnGDP.png")
```

Figure 8 shows a strong linear relationship between (log) NTL and GDP. In the year 2014, NTL explained 89% of the regional variation in GDP. Over time, the predictive power of NTL has remained stable around 90%. The regression coefficient of NTL in 2014 was 0.82, indicating that a 10% increase in NTL is associated with an 8.2% increase in GDP. Over time, this coefficient has slightly increased. By 2019, a 10% increase in NTL is associated with an 8.5% increase in GDP. Taken together, these results indicate that, on a year-by-year basis, nightlights are a useful proxy for economic activity.

5.2 Predicting GDP with nightlights

To evaluate the usefulness of nighttime lights (NTL) for predicting economic activity (GDP), let us consider the following panel-data model:



Note: Find the interactive version of this graph at Colab

Figure 8: Relationship between nighttime lights and GDP

$$\log(GDP)_{it} = \beta \log(NTL)_{it} + \mu_i + \varphi_t + \varepsilon_{it}, \tag{1}$$

where i indexes the regional economies, t indexes the years, μ_i is a region-specific effect, φ_t is a year-specific effect, and ε_{it} is random disturbance. Region-specific effects, μ_i, capture the influence of unobserved factors that are constant over time. Time specific effects, φ_t, capture the influence of unobserved factors that change over time but are common between regions. The most important parameter in this model is β, which summarizes the relation between GDP and nighttime lights (NTL). Given the logarithmic specification of the model, the parameter β indicates by what percentage GDP changes when the NTL changes by 1%. However, the specification of Equation 1 does not imply that NTL causes GDP. The parameter β only has a predictive interpretation.

There are multiple ways to estimate the parameter β. Let us consider the following three basic cases:

$$\log(GDP)_{it} = \beta_{\text{Pooled}} \log(NTL)_{it} + \mu + \varepsilon_{it}, \tag{2}$$

$$\overline{\log(GDP)_i} = \beta_{\text{Between}} \overline{\log(NTL)_i} + \mu_i + \overline{\varepsilon_i}, \tag{3}$$

$$\log(GDP)_{it} - \overline{\log(GDP)_i} = \beta_{\text{Within}} \left[\log(NTL)_{it} - \overline{\log(NTL)_i} \right] + \varphi_t + \varepsilon_{it} - \overline{\varepsilon_i}, \tag{4}$$

The simplest estimation of β is based on the so-called “pooled” estimator, β_{Pooled}. In this setting (Equation 2), time-specific effects are set to zero and all regions share a common intercept μ. The parameter β_{Pooled} indicates that–for all regional observations–an increase in NTL of 1% leads to a β_{Pooled} % expected increase in GDP. This model implies that we can expect the same effect of NTL on GDP if there is a 1% difference between regions or a 1% increase within a region. Thus, an important limitation of Equation 2 is that we cannot disentangle the usefulness of NTL data to predict cross-sectional differences or time series changes in GDP.

The “between” and “within” estimators are commonly used to evaluate the usefulness of NTL data for predicting GDP differences and changes within regions, respectively (Gibson et al. 2021, Zhang, Gibson 2022). In Equation 3, the (log) values of GDP and NTL are time averaged, and the model is estimated using standard cross-sectional methods. The parameter β_{Between} indicates the effect on GDP when NTL changes between regions. In Equation 4, Equation 3 is subtracted from Equation 1, and, by doing so, unobservable region-specific effects (μ_i) are removed from the estimation. The parameter β_{Within} indicates the effect on GDP when NTL changes within regions.

Table 7: The relationship between NTL and GDP

		Model Comparison	
	(1) Pooled	(2) Between	(3) Within
Dep. Variable	lnGDP	lnGDP	lnGDP
Estimator	PooledOLS	BetweenOLS	PanelOLS
No. Observations	216	36	216
Cov. Est.	Clustered	Clustered	Clustered
R-squared	0.8981	0.9036	0.0002
R-Squared (Within)	-0.2222	-0.2256	-0.0119
R-Squared (Between)	0.9036	0.9036	-0.0315
R-Squared (Overall)	0.8981	0.8980	-0.0314
F-statistic	1885.1	318.67	0.0381
P-value (F-stat)	0.0000	0.0000	0.8454
Intercept	8.6348	8.6059	18.260
	(27.551)	(11.013)	(19.150)
lnNTL	0.8433	0.8459	-0.0146
	(32.761)	(13.153)	(-0.1717)
Effects			Entity Time

T-stats reported in parentheses

The `Linearmodels` package allows us to estimate a variety of panel-data models. With this package, we can easily compare the previously described estimation approaches. Consistent with the previous literature (Gibson et al. 2021, Zhang, Gibson 2022), the results of Table 7 show that the predictive capabilities of NTL vary greatly depending on the type of data structure. The results of the "between estimator" are encouraging in terms of statistical significance and predictive power. NTL data predict about 90% of the variation in GDP data. The regression coefficient indicates that a 10% increase in NTL is associated with an 8.5% increase in GDP. In contrast, the results of the "within estimator" do not show a statistically significant relationship between NTL and GDP. Based on these results, nightlights perform much better in predicting cross-sectional differences than changes over time.

[42]:

```python
# @title CODE: Read panel dataset

df_panel = pd.read_csv("data/tabular/df_NTL_GDP_lnNTL_lnGDP.csv").set_index(
    ["region", "year"]
)
```

[43]:

```python
# @title CODE: Conduct panel data regressions

table = {
    "(1) Pooled": PooledOLS.from_formula(
        formula="lnGDP ~ 1 + lnNTL", data=df_panel
    ).fit(cov_type="clustered"),
    "(2) Between": BetweenOLS.from_formula(
        formula="lnGDP ~ 1 + lnNTL", data=df_panel
    ).fit(cov_type="clustered"),
    "(3) Within": PanelOLS.from_formula(
        formula="lnGDP ~ 1 + lnNTL + EntityEffects + TimeEffects", data=df_panel
    ).fit(cov_type="clustered"),
}
```

[44]:

```python
# @title CODE: Show comparative regression table

compare(table).summary
```

5.3 Comparing regional inequality dynamics: GDP vs nightlights

Inter-regional inequality is commonly identified as an important driver of socioeconomic instabilities, civil unrest, and political polarization (Ezcurra 2019, Rodríguez-Pose 2018). As a proxy of economic activity, NTL data are also used to understand regional inequality and its dynamics (Lessmann, Seidel 2017, Mendez, Santos-Marquez 2021, Mveyange 2018). In this section, we compare the evolution of regional inequality through the lens of GDP and NTL. For this purpose, we use two well-known inequality indicators: the Gini index and the Theil index.

The `inequality` package allows us to estimate both the Gini and Theil indexes. As we want to measure regional inequality for each year, we first need to define a function that computes regional inequality for each year within our dataset. Before applying these functions, we need to define a string-type vector that contains the time horizon of the analysis. The following code accomplishes these tasks.

[45]:

```python
# @title CODE: Define function: Gini index by column

def gini_by_col(column):
    return inequality.gini.Gini(column.values).g
```

[46]:

```python
# @title CODE: Define function: Theil index by column

def theil_by_col(column):
    return inequality.theil.Theil(column.values).T
```

[47]:

```python
# @title CODE: Define time index

years = np.arange(START_YEAR, END_YEAR + 1).astype(str)
```

Next, we apply these functions to the wide-form datasets: `gdf_lnGDP` and `gdf_lnNTL`. Four new datasets are created: `gini_lnGDP`, `gini_lnNTL`, `theil_lnGDP`, `theil_lnNTL`. The content and layout structure of these data can easily allow for further data processing or visualization.

[48]:

```python
# @title CODE: Calculate Gini index by column

gini_lnGDP = gdf_lnGDP[years].apply(gini_by_col, axis=0).to_frame("Gini_lnGDP")
gini_lnNTL = gdf_lnNTL[years].apply(gini_by_col, axis=0).to_frame("Gini_lnNTL")
```

[49]:

```python
# @title CODE: Calculate Theil index by column

theil_lnGDP = gdf_lnGDP[years].apply(theil_by_col, axis=0).to_frame("Theil_lnGDP")
theil_lnNTL = gdf_lnNTL[years].apply(theil_by_col, axis=0).to_frame("Theil_lnNTL")
```

Figures 9 and 10 provide a comparative visualization of the evolution of regional inequality, both in terms of nightlight luminosity (NTL) and economic activity (GDP). This comparative analysis indicates that, when measured against the regional disparities in GDP, the luminosity patterns reveal a higher degree of regional inequality. Specifically, Figure 10 shows that the inequality in NTL was about 1.83 times higher than the inequality in GDP in 2014. By 2019, this inequality ratio has been reduced to just above 1.76 times. From the perspective of the Theil inequality index, Figures 11 and 12 also indicate that the regional inequality in luminosity is higher than the regional inequality in GDP.

Researchers should be careful when interpreting the differences between NTL and GDP. Both types of data are subject to measurement errors. In particular, in the context of developing countries, GDP data can suffer from incomplete coverage, price distortions, and political distortions. Even higher-quality nightlights can suffer from ephemeral lights, background noise, and other measurement errors. Therefore, understanding the magnitude of these errors is crucial in drawing conclusions about the disparities in regional economic activity.

[50]:

```python
# @title CODE: Plot Gini index dynamics of (ln) GDP and NTL

df_gini = pd.merge(gini_lnGDP, gini_lnNTL, left_index=True, right_index=True)
df_gini.plot()
plt.ylabel("Gini index")
plt.savefig("figures/fig_ts_gini.png")
```

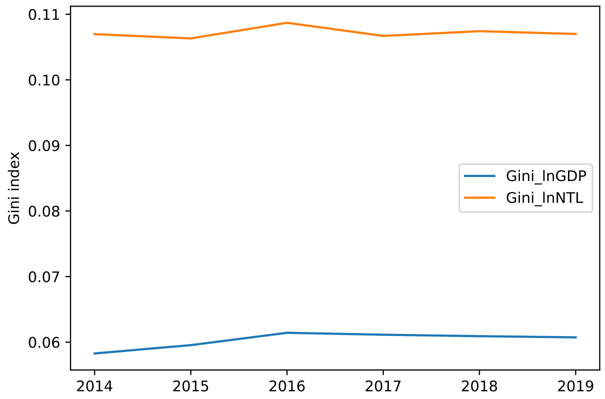


Figure 9: Regional inequality dynamics of GDP and NTL based on the Gini index

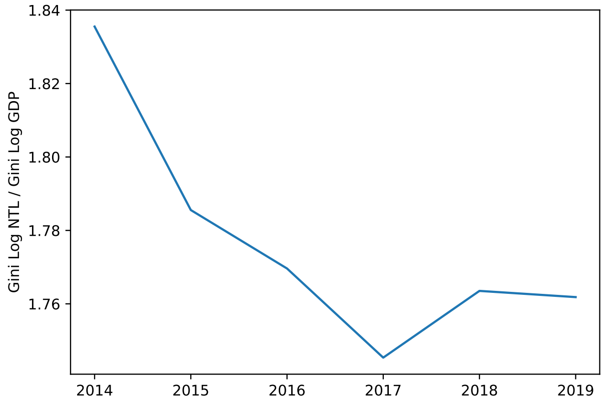


Figure 10: Gini-based inequality ratio between NTL and GDP

[51]:

```
# @title CODE: Plot inequality ratio (NTL/GDP) based on the Gini index

df_gini["Gini_Ratio"] = df_gini["Gini_lnNTL"] / df_gini["Gini_lnGDP"]
df_gini["Gini_Ratio"].plot()
plt.ylabel("Gini Log NTL / Gini Log GDP")
plt.savefig("figures/fig_ts_giniRatio.png")
```

[52]:

```
# @title CODE: Plot Theil index dynamics of (ln) GDP and NTL

df_theil = pd.merge(theil_lnGDP, theil_lnNTL, left_index=True, right_index=True)
df_theil.plot()
plt.ylabel("Theil index")
plt.savefig("figures/fig_ts_theil.png")
```

[53]:

```
# @title CODE: Plot inequality ratio (NTL/GDP) based on the Gini index

df_theil["Theil_Ratio"] = df_theil["Theil_lnNTL"] / df_theil["Theil_lnGDP"]
df_theil["Theil_Ratio"].plot()
plt.ylabel("Theil Log NTL / Theil Log GDP")
plt.savefig("figures/fig_ts_theilRatio.png")
```

6 Concluding remarks

The increasing availability of satellite nighttime lights can foster the monitoring of economic activity, especially in countries with limited official statistics. Luminosity at night is positively correlated with GDP and other economic measures across countries and subnational regions. However, satellite images can still present challenges that affect

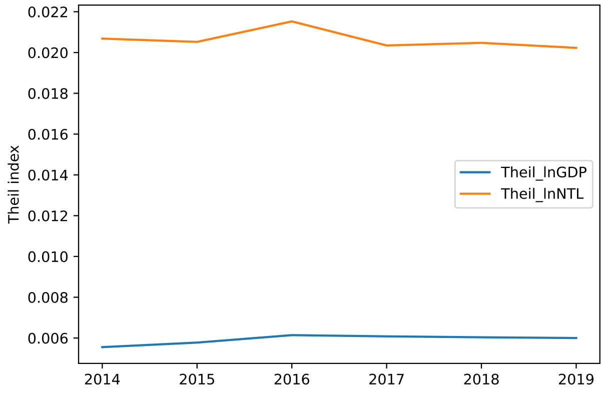


Figure 11: Regional inequality dynamics of GDP and NTL based on the Theil index

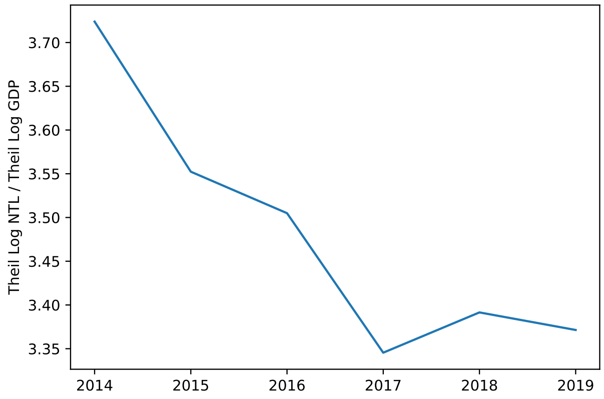


Figure 12: Theil-based inequality ratio between NTL and GDP

their application. A careful researcher must be aware of the magnitude of noise and measurement errors inherent in this kind of data.

Measurement errors in NTL data can arise from multiple sources, including ephemeral lights, sensor calibration issues, blurring effects, thresholding, and angular variations in satellite detection. Despite these concerns, there is a rapidly expanding body of literature that offers improved data products and novel tools to use these images effectively. This notebook introduced some of those data and tools to encourage further exploration and discussion of the measurement of regional economic activity.

In this notebook, we presented a user-friendly approach for analyzing satellite NTL images in a cloud-based Python environment. When using these data, one needs to interactively explore space-time patterns, as NTL may require additional cleaning and processing. In particular, when using NTL data to predict economic activity, one must note the difference between cross-sectional and time-series predictions. NTL data has been shown to perform much better with the former. Another application worth exploring is the measurement of regional inequality dynamics. Using multiple inequality measures is recommended to confirm regional inequality trends.

Links to the computational notebook

- Short link: https://bit.ly/project2022p
- Full link: https://colab.research.google.com/github/quarcs-lab/project2022p/blob/-master/project2022p_notebook.ipynb
- Github repository: https://github.com/quarcs-lab/project2022p

References

Abrahams A, Oram C, Lozano-Gracia N (2018) Deblurring DMSP nighttime lights: A new method using Gaussian filters and frequencies of illumination. *Remote Sensing of Environment* 210: 242–258. CrossRef

Alesina A, Michalopoulos S, Papaioannou E (2016) Ethnic inequality. *Journal of Political Economy* 124: 428–488. CrossRef

Bresenham JE (1965) Algorithm for computer control of a digital plotter. *IBM Systems Journal* 4: 25–30. CrossRef

Chen M, Fahrner D, Arribas-Bel D, Rowe F (2020) A reproducible notebook to acquire, process and analyse satellite imagery. *REGION* 7: R15–R46. CrossRef

Chen X, Nordhaus WD (2011) Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences* 108: 8589–8594. CrossRef

Donaldson D, Storeygard A (2016) The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives* 30: 171–198. CrossRef

Elvidge CD, Baugh K, Zhizhin M, Hsu FC, Ghosh T (2017) VIIRS night-time lights. *International Journal of Remote Sensing* 38: 5860–5879. CrossRef

Elvidge CD, Baugh KE, Zhizhin M, Hsu FC (2013) Why VIIRS data are superior to DMSP for mapping nighttime lights. *Proceedings of the Asia-Pacific Advanced Network* 35: 62. CrossRef

Elvidge CD, Zhizhin M, Ghosh T, Hsu FC, Taneja J (2021) Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing* 13: 922. CrossRef

Ezcurra R (2019) Interregional inequality and civil conflict: Are spatial disparities a threat to stability and peace? *Defence and Peace Economics* 30: 759–782. CrossRef

Falchetta G (2023) blackmarble: retrieve, wrangle and plot VIIRS Black Marble nighttimelight data in R. Github repository

Gibson J (2020) Better night lights data, for longer. *Oxford Bulletin of Economics and Statistics* 83: 770–791. CrossRef

Gibson J, Olivia S, Boe-Gibson G, Li C (2021) Which night lights data should we use in economics, and where? *Journal of Development Economics* 149: 102602. CrossRef

Gibson J, Olivia S, Boe-Gibson G (2020) Night lights in economics: Sources and uses. *Journal of Economic Surveys* 34: 955–980. CrossRef

Henderson JV, Storeygard A, Weil DN (2012) Measuring economic growth from outer space. *American Economic Review* 102: 994–1028. CrossRef

Kim B, Gibson J, Boe-Gibson G (2024) Measurement errors in popular night lights data may bias estimated impacts of economic sanctions: Evidence from closing the Kaesong Industrial Zone. *Economic Inquiry* 62: 375–389. CrossRef

Lessmann C, Seidel A (2017) Regional inequality, convergence, and its determinants – A view from outer space. *European Economic Review* 92: 110–132. CrossRef

Levin N, Kyba CC, Zhang Q, Sánchez de Miguel A, Román MO, Li X, Portnov BA, Molthan AL, Jechow A, Miller SD, Wang Z, Shrestha RM, Elvidge CD (2020) Remote sensing of night lights: A review and an outlook for the future. *Remote Sensing of Environment* 237: 111443. CrossRef

Li X, Zhou Y, Zhao M, Zhao X (2020) A harmonized global nighttime light dataset 1992–2018. *Scientific Data* 7. CrossRef

Li X, Zhou Y, Zhao M, Zhao X (2021) Harmonization of DMSP and VIIRS nighttime light data from 1992-2020 at the global scale. Figshare. dataset

Mendez C, Santos-Marquez F (2021) Regional convergence and spatial dependence across subnational regions of ASEAN: Evidence from satellite nighttime light data. *Regional Science Policy and Practice* 13: 1750–1777. CrossRef

Miethe J (2023) Nightlightstats: R-package. Github repository

Mveyange A (2018) *Measuring and Explaining Patterns of Spatial Income Inequality from Outer Space: Evidence from Africa*. World Bank, Washington, DC. CrossRef

Njuguna C (2020) Rnightlights: R package to extract data from satellite nightlights. Github repository

Patnaik A, Shah A, Thomas S (2023) Nighttimelights: Package to analyse VIIRS nighttime lights. Github repository

Raschky P (2020) nighttimelights: A repository of python scripts to calculate various nighttime light statistics. Github repository

Reades J (2020) Teaching on Jupyter. *REGION* 7: 21–34. CrossRef

Rodríguez-Pose A (2018) The revenge of the places that don't matter (and what to do about it). *Cambridge Journal of Regions, Economy and Society* 11: 189–209. CrossRef

Román MO, Wang Z, Sun Q, Kalb V, Miller SD, Molthan A, Schultz L, Bell J, Stokes EC, Pandey B, Seto KC, Hall D, Oda T, Wolfe RE, Lin G, Golpayegani N, Devadiga S, Davidson C, Sarkar S, Praderas C, Schmaltz J, Boller R, Stevens J, Ramos González OM, Padilla E, Alonso J, Detrés Y, Armstrong R, Miranda I, Conte Y, Marrero N, MacManus K, Esch T, Masuoka EJ (2018) NASA's black marble nighttime lights product suite. *Remote Sensing of Environment* 210: 113–143. CrossRef

Rowe F, Maier G, Arribas-Bel D, Rey S (2020) The potential of notebooks for scientific publication, reproducibility and dissemination. *REGION* 7: E1–E5. CrossRef

Smits J, Permanyer I (2019) The subnational human development database. *Scientific Data* 6. CrossRef

Sutton P, Elvidge C, Ghosh T et al. (2007) Estimation of gross domestic product at sub-national scales using nighttime satellite imagery. *International Journal of Ecological Economics and Statistics* 8: 5–21

Zhang X, Gibson J (2022) Using multi-source nighttime lights data to proxy for county-level economic activity in China from 2012 to 2019. *Remote Sensing* 14: 1282. CrossRef

Zhang X, Gibson J, Deng X (2023) Remotely too equal: Popular DMSP night-time lights data understate spatial inequality. *Regional Science Policy and Practice* 15: 2106–2125. CrossRef

Zheng Q, Seto KC, Zhou Y, You S, Weng Q (2023) Nighttime light remote sensing for urban applications: Progress, challenges, and prospects. *ISPRS Journal of Photogrammetry and Remote Sensing* 202: 125–141. CrossRef

Appendices

A. Zonal statistics

A.1 Main logic

Let us consider the code inside this loop:

[54]:
```
for year in range(START_YEAR, END_YEAR + 1):
    # The raster file of the given year is loaded
    raster_file = load_raster(year)

    # The mask is applied, and then the aggregator operation is performed for computing
    # the aggregate radiance.
    statewise_agg_ntl = polygons_files.geometry.apply(geom_mask,
                        dataset=raster_file).apply(AGGREGATE_OPERATOR)

    # The state-wise aggregate radiance of the year is stored in the data frame that
    # was initialized earlier.
    gdf_NTL[str(year)] = statewise_agg_ntl
```

In particular, let us evaluate last expression:

[55]:
```
polygons_files.geometry.apply(geom_mask, dataset=raster_file).apply(AGGREGATE_OPERATOR)
```

The `apply` method is used to apply a function on all elements of a column of a data frame. This column contains the polygons of states of India. First, we want to apply the `mask` function from `rasterio` to return a matrix corresponding to a raster file, which has `nodata` at all locations outside the polygons of the states.

[56]:
```
def geom_mask(geom, dataset, crop=True, all_touched=True):
    masked, mask_transform = mask(dataset=dataset, shapes=(geom,), crop=crop,
                             all_touched=all_touched)
    return masked
```

For example:

[57]:
```
geom_mask(polygons_files.geometry[0])
```

Applies to the first polygon and returns a matrix with `nodata` outside the first polygon; hence

[58]:
```
polygons_files.geometry.apply(geom_mask)
```

Applies to all polygons.

Second, after attaining the list of matrices with `nodata` outside the polygons, we want to apply the `AGGREGATOR_OPERATOR` to get the aggregate the light inside the polygons.

For example:

[59]:
```
AGGREGATE_OPERATOR(geom_mask(polygons_files.geometry[0]))
```

Since `AGGREGATE_OPERATOR = np.ma.sum` (by default), this returns the sum of light of the first polygon, hence

[60]:
```
polygons_files.geometry.apply(geom_mask).apply(AGGREGATE_OPERATOR)
```

Returns the sum of light of all polygons.

A.2 geom_mask function

We want to apply the `mask` function from `rasterio` on all elements of a `pandas` data frame column using `apply`, however, we need to pass additional arguments. For this, we create a wrapper function called `geom_mask`, which calls `mask` and passes the additional arguments. Additionally, `mask` returns two values, we only need the first one, hence `geom_mask` is used to select only one.

A.2.1 Positional arguments

The function has two positional arguments; geom and dataset.

A.2.2 Keyword arguments

There are two keyword arguments, `crop` and `all_touched`.

The `crop = True` is essential. It is used to crop the output matrix to the extent of the polygon. This substantially reduced the memory requirement of the program.

The second keyword argument, `all_touched = True` tells the mask function to include pixels which are touching boundaries. If false, the function will include a pixel only if its center is within the boundaries, or if it is selected by Bresenham (1965) line algorithm. For most polygons in our case, `all_touched = False` will produce similar results. The difference would be noticeable for states that contain islands.

A.3 For loop

The iterator in the loop is used for two reasons. First, it is used to open the raster file corresponding to a year.

[61]:
```
raster_file = load_raster(year)
```

Second, it is used to create a column to store the state-wise sum of lights

[62]:
```
gdf_NTL[str(year)] = statewise_agg_ntl
```

Hence, for each raster file, the state-wise sum of lights is computed and stored in a column named year.

B. Folder structure

The notebook's root directory is organized into three primary folders: `data`, `figures`, and `tables`. These folders are created during the execution of the notebook.

- **data:** Contains all input data that is downloaded during the execution of the notebook.
  - *raster:* Contains the VIIRS nighttime lights raster files.
  - *tabular:* Contains state-level GDP data for India (`df_GDP_India36.csv`), used as ground truth. During processing, the file `df_NTL_GDP_lnNTL_lnGDP.csv` is saved in this directory. It contains a dataframe with logs of aggregate nighttime lights and GDP of each state from the start year to end year.
  - *vector:* Contains the geojson files. `gdf_india36.geojson`, which is downloaded, contains the shape of each state of India. During processing, the following files are saved: `gdf_lnGDP.geojson`, `gdf_lnNTL.geojson`, and `gdf_NTL.geojson`. These contain state-wise log GDP, log aggregate nighttime lights, and aggregate nighttime lights, respectively.
- **figures:** Stores all figures generated during notebook processing.
- **tables:** Stores all tables generated during notebook processing.

C. Harmonized nighttime lights

Li et al. (2020) have constructed a harmonized annual nighttime lights dataset by combining the DMSP-OLS and the VIIRS datasets. This newly extended data set is useful because it provides a long-term series of data that is often required to study long-term changes in economic activity. Harmonized nighttime lights from 1992 to 2021 are available from the following figshare repository: https://doi.org/10.6084/m9.figshare.9828827.v7. The authors have made data downloading easy and accessible. The following code block downloads and extracts the images.

[63]:
```
from io import BytesIO # This in a standard library.

# Define the URL of the zip file
```

```
url = "https://figshare.com/ndownloader/articles/9828827/versions/7"

# Send an HTTP GET request to the URL
response = requests.get(url)

# Check if the request was successful (status code 200)
if response.status_code == 200:
    # Read the content of the response as bytes
    zip_content = BytesIO(response.content)

    # Extract the zip file
    with zipfile.ZipFile(zip_content, 'r') as zip_ref:
        # You can list the files in the zip file if needed
        zip_ref.printdir()

        # Extract the files to a directory (e.g., 'my_extracted_files/')
        zip_ref.extractall(RASTER_DIRECTORY)

    print("Zip file has been successfully extracted.")
else:
    print("Failed to retrieve the zip file. Check the URL and your internet connection.")
```

The `load_raster` function in the notebook can be replaced by the following for easily loading the images.

[64]:

```
def load_raster(year):
    """
    Load a raster file based on the provided time identifier.

    Parameters:
    -----------
    year : int

    Returns:
    --------
    rasterio.io.DatasetReader
    An opened raster file dataset ready for further operations.

    Example:
    --------
    >>> raster_2014 = load_raster(2014)
    >>> type(raster_2014)
    <class 'rasterio.io.DatasetReader'>

    Notes:
    ------
    Modify the path in the function if your file structure
    or naming convention differs.
    """
    raster_path = f"{RASTER_DIRECTORY}/Harmonized_DN_NTL_{year}*"
    return (rasterio.open(glob.glob(raster_path)[0]))
```

By changing the downloading and loading code, one can use the harmonized nighttime lights dataset without other changes.

Note: When using this dataset, keep radiance threshold as 63 in Figure 1.

D. Export tables in TeX

[65]:

```
# @title CODE: Define functions

def save_latex_table(df, filepath, max_rows=None, max_cols=None):
    truncated_df = pd.read_html(
        df.copy().round(4).to_html(index=False, max_rows=max_rows, max_cols=max_cols)
    )[0]
    print(truncated_df.astype(str).to_latex(index=False), file=open(filepath, "w"))
```

[66]:

```
# @title CODE: Save abbreviated zonal statistics table

save_latex_table(gdf_NTL, "tables/gdf_NTL.tex", 10, 7)
```

[67]:

```
# @title CODE: Save abbreviated long-form dataset as TeX file

save_latex_table(df, "tables/df.tex", 10, 7)
```

[68]:

```
# @title CODE: Save abbreviated geospatial dataset as TeX table

save_latex_table(gdf_lnNTL, "tables/gdf_lnNTL.tex", 10, 7)
```

[69]:

```
# @title CODE: Save abbreviated geospatial dataset of (ln) NTL as TeX table

save_latex_table(gdf_lnNTL, "tables/gdf_lnGDP.tex", 10, 7)
```

[70]:

```
# @title CODE: Save table in .tex format

print(compare(table).summary.as_latex(), file=open("tables/panel_regression.tex", "w"))
```

[71]:

```
# @title CODE: Save the summary statistics table

print(gdf_NTL_summary.to_latex(), file=open("tables/gdf_NTL_round.tex", "w"))
```

Funded by

ersa

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

FWF