Regional development at the NUTS 3 level in Europe between 2010 and 2021. What sorts of regions are there, where are they, and how can we use that information to inform evidence based regional development policy?

Dr Becky Arnold, Ewoud T. Jansma MSc, and Prof. Leo van Wissen

July 30, 2024

Abstract

The Regional Development Characteristics Database, introduced and described in [3], is presented and analysed. This database, compiled as part on the Horizon Europe project PREMIUM_EU, collates 53 regional development indicators spanning from GDP per capita, to air pollution, to crime rates. These indicators are divided into economic, social, living environment, political, and geographic dimensions, and are collected at the NUTS 3 level for European countries between 2010 and 2021. The resulting database, while an extremely rich and powerful resource, sufferers from significant incompleteness and inhomogeneity. In this paper these issues are mitigated, by performing a disaggregation, which draws on data available at larger geographical scales (i.e. NUTS 2) to model missing data at the NUTS 3 level.

The database is analysed, and its potential for examining a variety of questions relating to regional development is demonstrated. The distribution of region's development in different dimensions is examined, and a strong linear correlation is found, e.g. a region that is underdeveloped in terms of living environment is typically underdeveloped in the economic and social dimensions too. From this a typology of regions is developed, dividing regions in vulnerable, underdeveloped, developed, and leading categories. The geographical distribution of regions as a function of their regional development is examined. On the macro scale a significant gradient is observed both from west to east and north to south, with more developed regions typically lying in the north and west. This is in line with expectations. Despite these macro trends, significant variation is observed at subnational scales. A surprisingly weak positive correlation between region's development and urbanisation is observed.

For each region the evolution of development score over time is then examined. There is a positive correlation between regional development and time in 768 regions, which we call improving regions. There is no significant correlation in 641 regions, referred to as stagnant regions. There is a negative correlation in 57 regions, termed declining regions. Inspection of the geographical distribution of these three groups shows that stagnant and declining regions are most commonly located in central Europe, parts of Scandinavia, Ireland, and in eastern Turkey. It is observed that improving regions disproportionately likely to be highly developed, and in contrast declining regions are disproportionately those that are already vulnerable or underdeveloped. This is concerning, as it forecasts escalating inequality.

The implications of these results for regional development policy at the European and national scales is then discussed, and the diversity of vulnerable and underdeveloped regions is highlighted. This diverse nature makes it likely that a wide range of policies will be necessary to confront each type of region's particular challenges. We note that this work is still in development and this draft paper should be considered a work in progress rather than a final scientific output.

1 Introduction

Regional development is one of the EU's principle strategies for increasing equality and living standards internationally. With a budget of € 226.05 billion for 2021 - 2027, the stated goal of the European Regional Development Fund is to 'reduce economic, social and territorial disparities' [1]. With the societal and financial stakes so high it is vitally important that the landscape of regional development within

Europe be systematically assessed. Only with a strong understanding of the nature and challenges of vulnerable regions can such efforts be put to the greatest benefit. Furthermore, regional development is not something linear or straightforward. It must also always be considered in the wider economic, political, and cultural context. That context itself spans from the NUTS 2 level all the way up to the global scale, with pressures and events at each level impacting upon one another.

Historically, conceptions of regional development have mostly focused on the economic domain, highlighting metrics such as GDP per capita, unemployment, and average educational levels. More recently, regional development has been regarded as a highly multifaceted concept which encapsulates social conditions, living environments, political contexts and demographics alongside those always-important economic factors [5].

These different dimensions of regional development must be coherently combined in order to understand:

- What makes regions vulnerable?
- What different types of vulnerable regions are there?
- What are the challenges that these different types of regions are facing, and what are the most efficient solutions to those challenges?

The Horizon Europe project, PREMIUM_EU aims to aid regional policy makers in identifying effective policies to develop their regions, with a particular focus on vulnerable regions and migration policy. The first step of this goal is to develop a typology of regions, and within this project we aim to do this at the NUTS 3 level. The progress of the project in this regard is described jointly by [3] and this paper.

The first paper in this sequence, [3], outlines the theoretical basis for our conceptualisation of regional development, and the indicators chosen to describe that conceptualisation. These indicators are wide ranging, spanning from the conventional, such as GDP per capita, to features such as air pollution, crime rates, and the number of doctors per 1000 people, among many others. They are collected together in the Regional Development Characteristics Database, presented in [3], which is divided in economic, social living environment, political, and geographic dimensions. Also introduced in that paper is the Regional Demographics Characteristics Database, but that is not the focus of this paper.

In this work the Regional Development Characteristics Database is analysed. In section 2 the data's completeness and geographic coverage is assessed and the steps taken to improve this completeness are outlined. The method used to convert this data into regional development is described. In section 3 the regional development typology we develop on the basis on this database is explained. A geographic and time series analysis in then performed, identifying how well/underdeveloped and improving/declining regions are distributed. In section 4 the policy implications of our findings are outlined, and in section 5 the future work we intend to carry out on and with this dataset is discussed.

2 Data

The Regional Characteristics Database contains 33 indicators which are used in this analysis. This data is primarily drawn from Eurostat and the OECD, with some supplementary data from sources such as the Global Data Lab [6]. The 33 indicators are divided into fifteen economic indicators, five living environment indicators, and thirteen social indicators. This database also contains eight geographic indicators, such as the area of the region, and whether the region is coastal. While these geographic indicators help define the profile of a region they are not inherently good or bad, so are not considered as regional *development* features, and are not included in this paper's analysis.

Also excluded from this analysis are the political indicators, of which there are ten. Because the PREMIUM_EU project is mostly focused on the impact of migration and migration policy in particular upon regions the policy indicators that have been collected are specifically selected to highlight that topic. Examples of political indicators in the database are the degree to which migrants enjoy the same rights as native citizens to participate politically, and their protections from discrimination. These indicators are excluded here because migration policy is almost always set at the national level, reducing the use of these indicators for assessing the development status of individual regions.

As noted in [3], the NUTS regional boundaries have gone through multiple iterations, and in this work only the NUTS 2021 boundaries are used. That means that regions that no longer exist, e.g. due to being merged into other regions, are excluded even if data has been collected for them. Newly created regions are included as far into the past as data has been collected for them. For some indicators data have been 'back calculated' by national statistical institutes. E.g. a new region is coined in 2021, and the relevant national statistical agency calculates what the GDP per capita was within those boundaries in previous years. As a result in some cases it is possible to produce regional development profiles for regions in years prior to their official creation.

Despite extensive work, and the deliberate bias of selected indicators towards those with high coverage, the regional characteristics database is far from complete, especially at the NUTS 3 level we wish to build regional profiles of. Overall, at NUTS 3 the economic indicators have an average completeness of 52.40 %. The living environment indicators have almost the same completeness, at 54.29 %. The social indicators have the worst completeness at 24.15 %.

Conventional statistical methods and machine learning are considered as methods to model the missing data and improve the completeness of the data set. Both are rejected because it is observed that the incompleteness of the data set is highly inhomogeneous. Data for a given indicator is often not collected in every year, or in every country. Even in countries where the data is collected different countries often collect the data on different NUTS levels (e.g. some collecting that data at NUTS 2 and others at NUTS 3). A country also may not collect the data in all the regions it contains, even at their chosen level. This inhomogeneity makes producing a meaningful training data set for machine learning or robust classical modelling extraordinarily challenging, and both approaches are determined to be non-viable (at least in the short term). One method that is used to increase the completeness of the data set is disaggregation, which is discussed in section 2.1, and improves the economic indicator's completeness to 84.74 %, the living environment indicator's completeness to 67.80 %, and the social indicator's completeness to 55.18 %

2.1 Disaggregation

In some cases data is available at the NUTS 2 level, but not at NUTS 3. While the true value at NUTS 3 remains unknown it is reasonable to assume it is highly correlated with that of its parent NUTS 2 region. In order to utilise this NUTS 2 data we disaggregate it to the NUTS 3 level. This helps reduce the number of empty cells in the regional development characteristics database.

Currently disaggregation is performed in the simplest possible way, which we term duplication. In this method we assume that all NUTS 3 regions within the NUTS 2 region have the same value of the property as is recorded at NUTS 2. An example of such a property is educational level (percentage of the population aged 25-64 with tertiary educational attainment). If this is X % in a NUTS 2 region, we assume it is also X % in the relevant NUTS 3 regions.

This is obviously a gross oversimplification, however we intend to improve on it in the future. One potential method under consideration is to use one or more indicators to define regions as 'core' or 'peripheral' (loosely translating to urban and rural), in order to perform a more nuanced analysis.

Another option that we intend to explore is looking at selections of indicators we suspect to be closely correlated, e.g. rates of intentional homicide, rates of assault, and the subjective feeling of safety of the population. By examining the relationships between these variables in regions where all three are collected we can estimate them in regions where only some of those properties are recorded.

2.2 Geographic completeness

The geographical coverage of each individual indicator was mapped and considered when the final list of indicators was determined. This was done to try to make a selection that was as geographically even-handed as possible, while also balancing data completeness and the indicator's theoretical merits, as outlined in [3]. Following this disaggregation the geographic coverage of all the indicators chosen is evaluated. This is done by producing colour coded plots, showing the number of indicators for which there is valid data in each NUTS 3 region. Fig. 1 shows the total number of indicators in each region in 2021. This is then split down into the economic (Fig. 2), living environment (Fig. 3), and social (Fig. 4) dimensions.

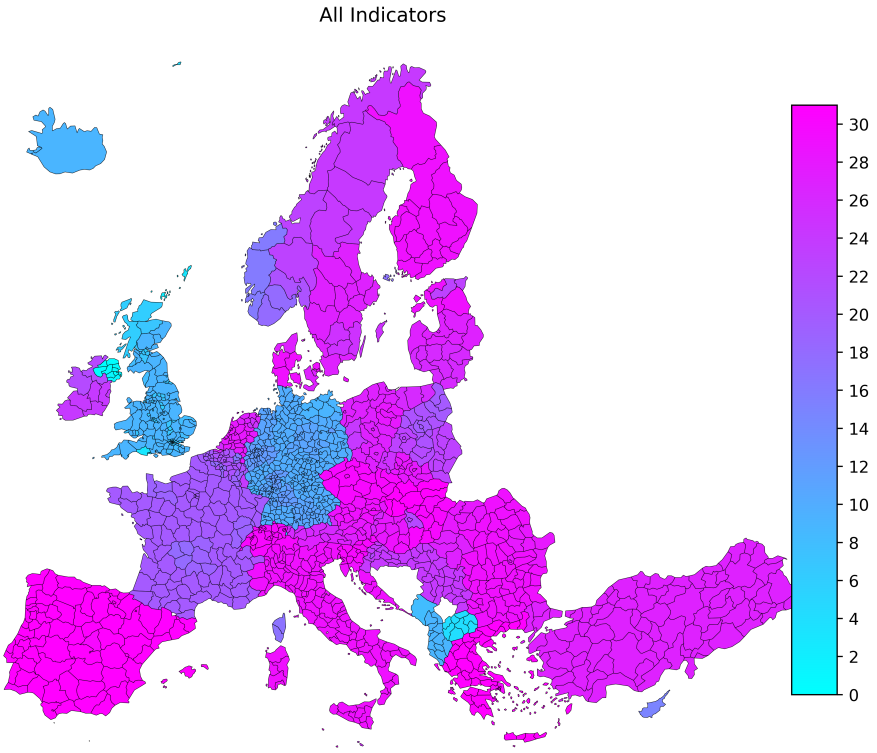

Figure 1: The number of all types of indicators recorded in each NUTS 3 region in 2021.

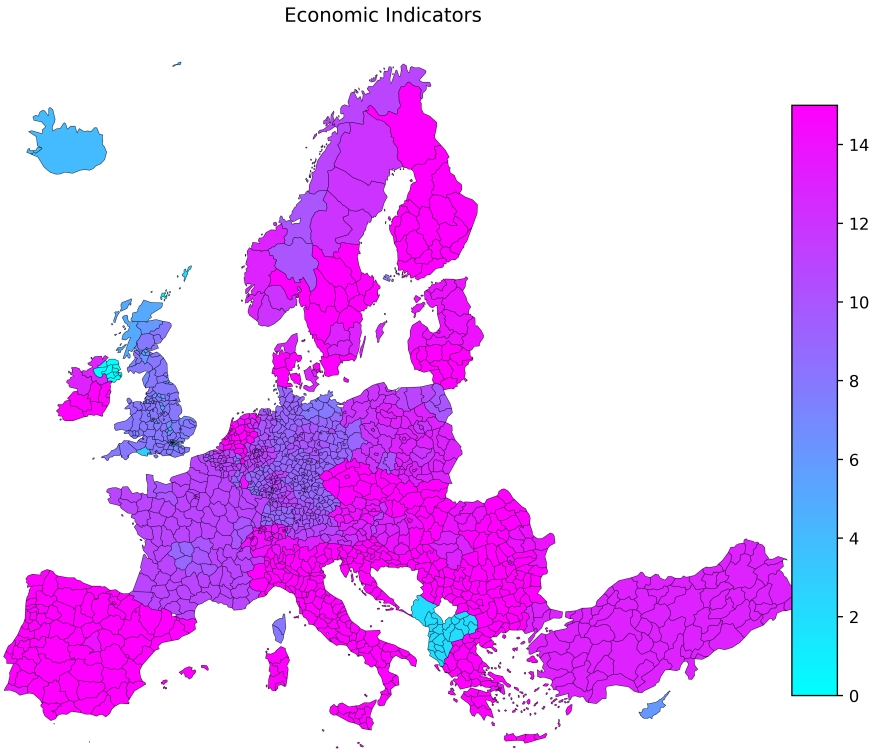

Figure 2: The number of economic indicators recorded in each NUTS 3 region in 2021.

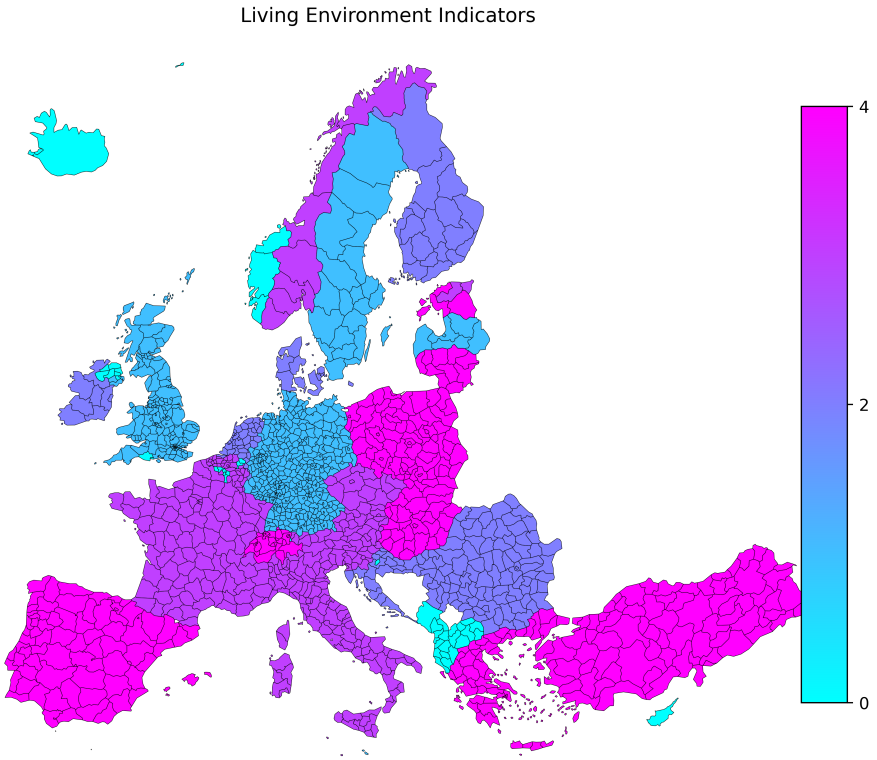

Figure 3: The number of living environment indicators recorded in each NUTS 3 region in 2021.

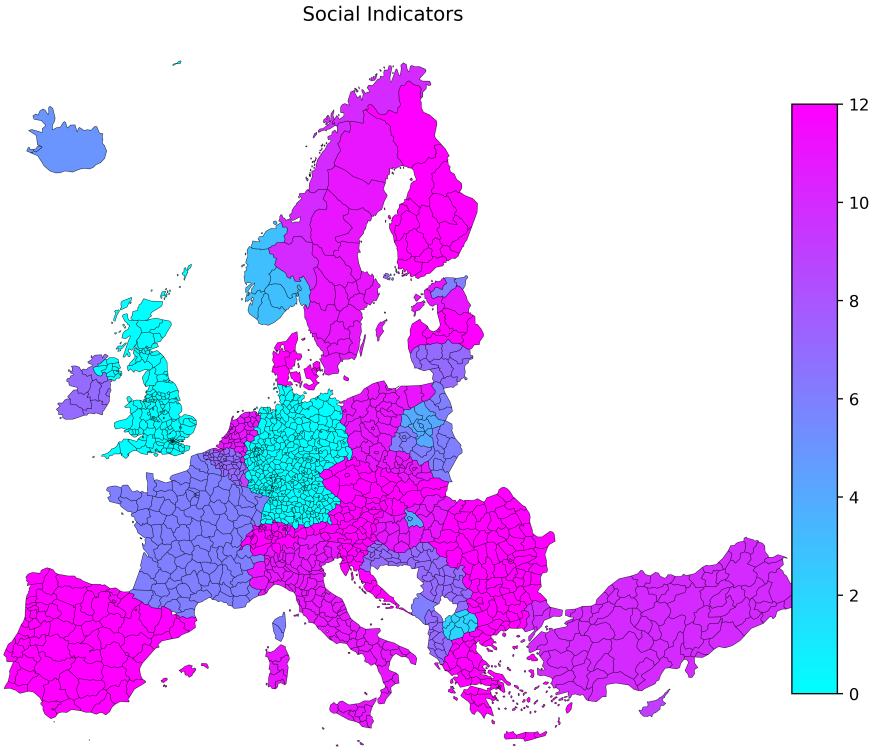

Figure 4: The number of social indicators recorded in each NUTS 3 region in 2021.

Table 1: An example of a data set with five regions, and two indicators, for which there has been incomplete and inhomogeneous data collection.

Region	Educational Level	GDP per Capita
A	72	40000
B	15	No data
C	No data	10000
D	No data	25000
E	43	5000

Table 2: The example data set where the values in each column have been replaced with ranks from best to worst.

Region	Educational Level	GDP per Capita
A	1	1
B	3	No data
C	No data	3
D	No data	2
E	2	4

2.3 Converting indicators into indices

Using the collected data, overall quality indices are calculated for each region for each year from 2010 to 2021 in each of the three dimensions (economic, social, living environment). The considerable sparsity and inhomogeneity of the data set presents a barrier to this, and so the method we have designed for this conversion is deliberately constructed to be resistant to that data weakness. The method for this will be described via a simplified example in which the economic dimension contains only two indicators, educational level and GDP per capita. To further simplify this example we consider data for only one year, and assume that there are only five regions, A, B, C, D, and E.

For the purposes of this explanation consider the example data set shown in Table 1.

Step 1: For each property rank the regions from best to worst. Note that if one of the properties was undesirable, e.g. unemployment rate, this ranking would done in reverse order with the smallest first and largest last. See Table 2 to see this reflected in the example data set.

Step 2: Convert those rankings to fractions normalised by the number of valid datapoints for each property. In the example case educational level has three valid entries, so ranking 1, 2, 3, becomes 1, 0.5, 0. In contrast GDP per capita has four valid entries, so ranks 1, 2, 3, 4 become 1, 0.66, 0.33, 0. The consequences of this step for the example data set can be seen in Table 3.

Step 3: For each region calculate the average of all its properties now they have been converted to fractions. This is the index, and for the example case can be seen in the last column of Table 3.

Step 4: Now this process has been used on the economic indicators to produce and economic index it is repeated for the social and living environment indicators to produce indices in each of those dimensions.

The resistance of this method to missing data is such that as long as our selection of indicators ensures every region has data for at least one property in each dimension then an index can be calculated. Despite

Table 3: The example data set where the ranks in each column have been replaced with fractions, and the average fraction in shown in the last column.

Region	Educational Level	GDP per Capita	Economic Index
A	1	1	1
B	0	No data	0
C	No data	0.33	0.33
D	No data	0.66	0.66
E	0.5	0	0.25

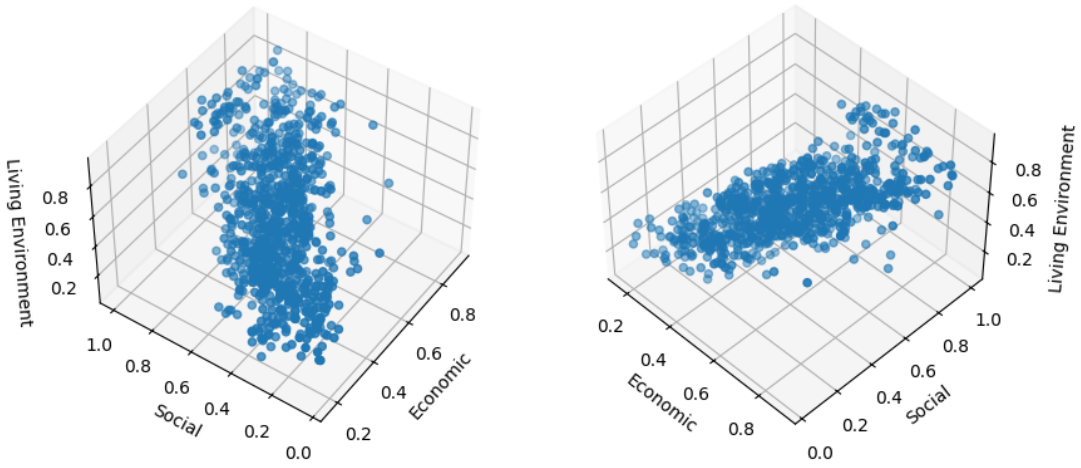

Figure 5: The regional development profiles for regions in 2021 plotted as points in the economic, social, and living environment dimensions.

this, due to the large number of individual NUTS 3 region-year combinations there are in the data set (18,168) there are some cases where no indicators in a certain dimension are available, and so cannot be computed. There are 23 such cases where the economic index can't be computed (0.13 %), 576 cases in the social dimension (3.17 %), and 1,171 cases in the living environment dimension (6.45 %). An example of this can be seen in Fig. 4, where it is visually apparent that none of the chosen social indicators were recorded in Germany or the UK in 2021 (though they were in other years).

Only region-year combinations where a valid value is computed in all three dimensions are included in the analysis presented in this paper[1]. N.B. By region-year combination we mean a particular NUTS 3 region and year, e.g. the region NL111 in 2020 has a different regional development profile to NL112 in 2020, and to NL111 in 2021. Thus the overall number of region-year combinations for which all three indices can be calculated (16,446) is the number of regional profiles in this data set.

3 Results

For each year the regions were plotted as points in three dimensional (economic, social, living environment) space. The results for one such year, 2021 (the most recent in the data set) is shown in Fig. 5. This figure shows regions are arranged in a fairly compact cylinder in this space, indicating that there is a strong correlation between region's development in all three dimensions. This is intuitive as we would expect regions that have, for example, low economic and social development to also have poor living environments.

We use this result to produce an overall development score for each region. To do this, first linear regression is used to fit a line in 3D space to the data set for each year. Next the mean gradient and intercept of all the lines is computed and averaged. This average line can be though of as an empirical axis of regional development which regions move up as they improve. Therefore we quantify how far along this axis a region lies as its overall regional development, the development index D. The coefficient of this line on the economic axis is 0.29 ± 0.04, the coefficient on the social axis is 0.71 ± 0.04, and the coefficient on the living environment axis is 0.65 ± 0.04. The zero point of this vector (where we define the zero point as where the economic development value is zero) is of course zero for the economic axis, -0.71 ± 0.21 in the social dimension and -0.57 ± 0.09 in the living environment dimension.

To compute D the position of each region along this 'axis of development' is computed. This is

[1] As a result there are no regional development profiles for Germany or the UK in 2021.

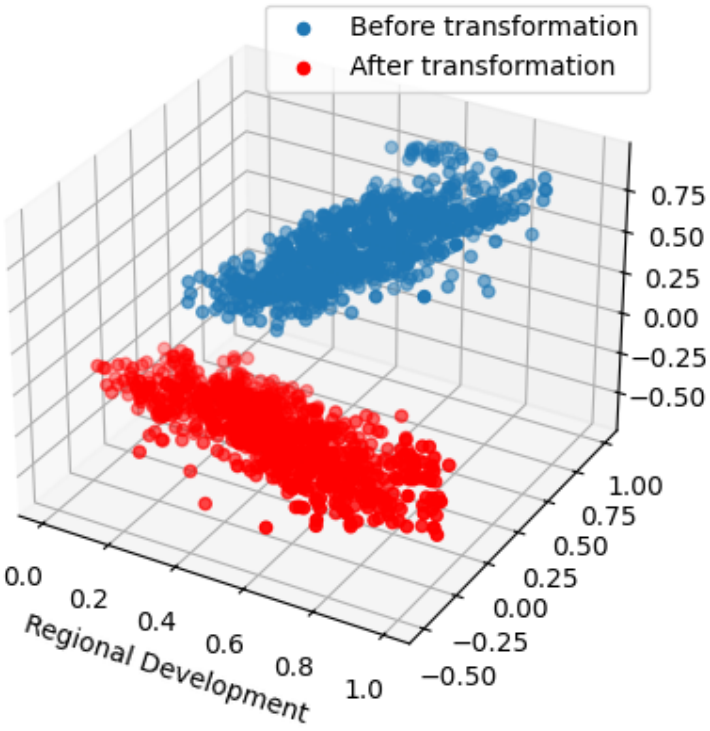

Figure 6: Figure showing the regional profiles in 2021, and transformation which makes them lie on the regional development axis.

done by rotating this axis to lie along the x axis, and all the datapoints are transformed along with it. The results are then rescaled such that the minimum x value is zero and the maximum is one. This transformation is shown for the 2021 data set in Fig 6. The x axis is now equivalent to the D axis, and a region's position along this axis is its regional development score.

The distribution of the regional development scores, D, is shown by a histogram in in Fig. 7. From this figure it is apparent that the distribution increases until it peaks at around $D = 0.4$. This is then followed by a sharp decline up to $D = 0.6$. Flowing this the distribution declines at a shallower rate until the maximum possible D is reached.

The regions are divided into four classes. Regions with $D < 0.25$ are categorised as vulnerable regions. Those with $0.25 < D < 0.5$ are underdeveloped regions. Regions where $0.5 < D < 0.75$ are referred to as developed regions. Finally when $D > 0.75$ the region is classified as a leading region. From Fig. 7 is apparent that regions are not evenly distributed between these categories. 1917 region-year combination are categorised as vulnerable, 7603 as underdeveloped, 4902 as developed, and 2024 as leading. If this list is limited to only regions in 2021 the totals become 133 vulnerable regions, 402 underdeveloped regions, 292 developed regions, and 68 leading regions. Fig. 8 shows the regions colour coded by their regional development score in 2016 (the mid point year of this data set). The results are broadly in line with expectations; western Europe shows a higher degree of development on average than eastern Europe (particularly Turkey), but there is considerable variation within countries. Despite this variation we note that even within countries there appears to be a high degree of spatial correlation in D, i.e. regions with relatively higher D tend to be in close proximity to regions that also have high D. Likewise, regions with relatively low D tend to be in close proximity to regions that also have low D. This observation is confirmed using Moran's I, which has an average value of 0.29 and a standard deviation of 0.02 across the years (2010-2021) in this data set.

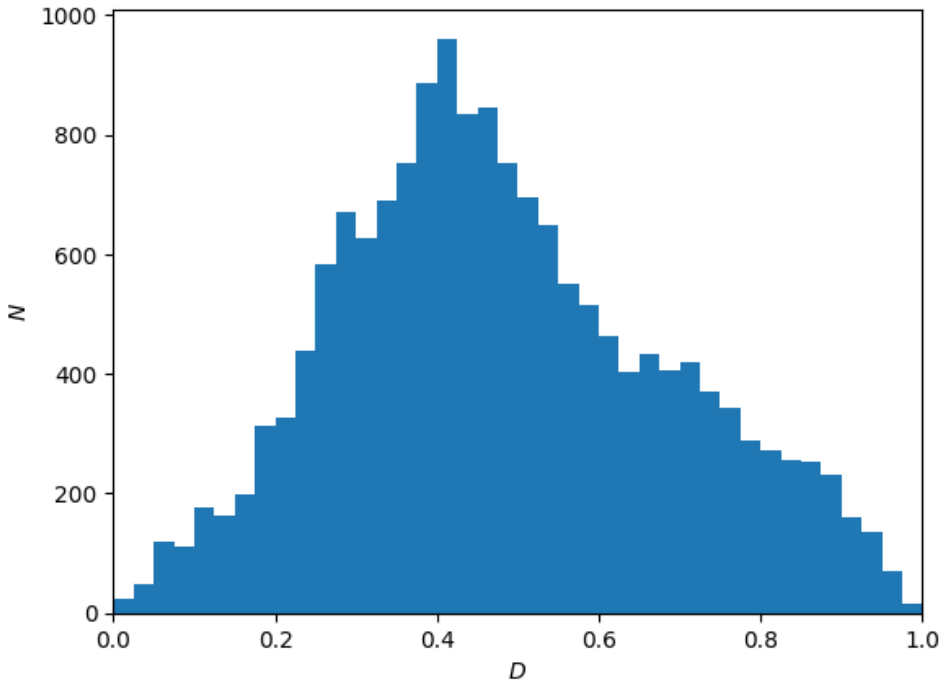

Figure 7: A histogram of the regional development scores for all region-year combinations in the database.

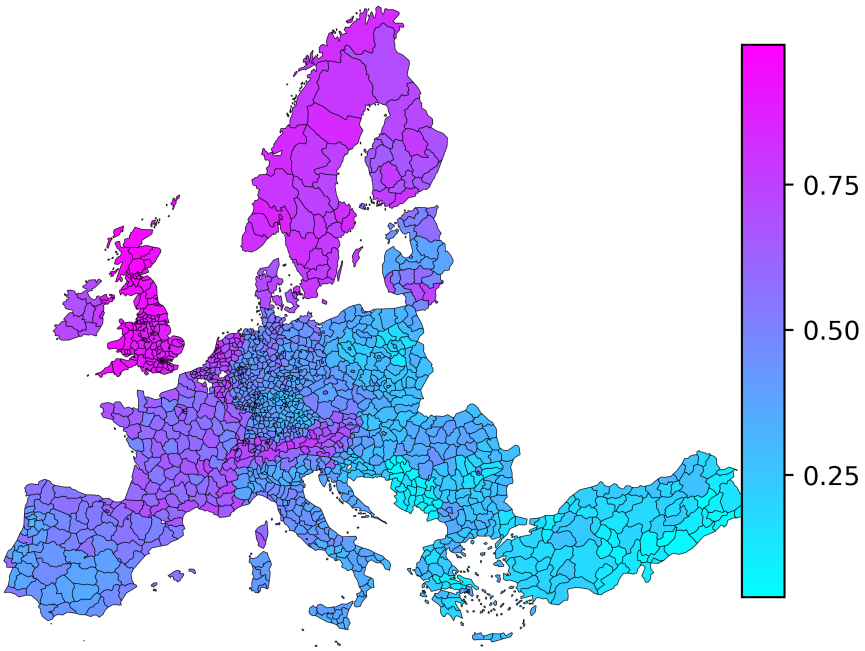

Figure 8: A map colour coded by regional development score, D, in 2016.

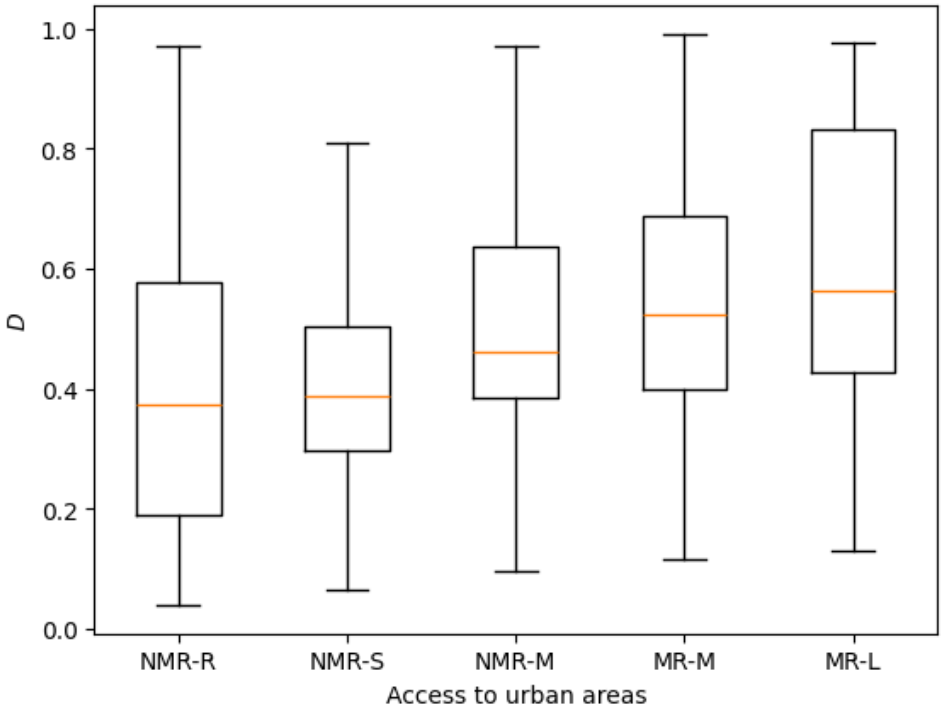

Figure 9: A box plot showing the median and interquartile range of regional development scores, D, for increasingly urban classifications of region. The data presented in this figure is for the midpoint year of the dataset, 2016.

3.1 Development and urbanisation

The relationship between development and geographic position is further explored by examining how D correlates with the region's access to functional urban areas, as defined and collated by the OECD's Regional Statistics Dataset [4]. In that data regions are divided into five categories, based on the work of [2]:

- NMR-R: a remote region
- NMR-S: near a small functional urban area of between 50,000 and 250,000 people
- NMR-M: near a midsize/large functional urban area with more than 250,000 people
- MR-M: a metropolitan midsize region between 250K and 1.5M inhabitants
- MR-L: a large metropolitan region with more than 1.5 million inhabitants.

In Fig. 9 these five region types are shown on the x axis in order of increasing urbanisation. The y axis shows regional development, D. This box plot shows the median and interquartile range of D values for regions of each urbanisation category. It is apparent from this figure that there is a shallow but convincing upwards trend in the median D (p-value < 0.01). This means that on average more urban areas are more developed than rural ones. That is largely in keeping in what would be naively expected, but is perhaps less obvious when the multidimensional aspect of the quantification of regional development used in this paper is considered. This conceptualisation, as outlined in the preceding paper, [3], and in sections 1 and 2 of this paper deliberately takes into account a wide range of indicators that go beyond the economic dimension to include metrics such as life satisfaction, crime rate, and more in the social and living environment dimensions.

This holistic approach may explain why, although there is a slight correlation between degree of urbanisation and regional development, all five region categories in Fig. 9 contain a diverse array of region development statuses. For each category the interquartile range spans the majority of the range

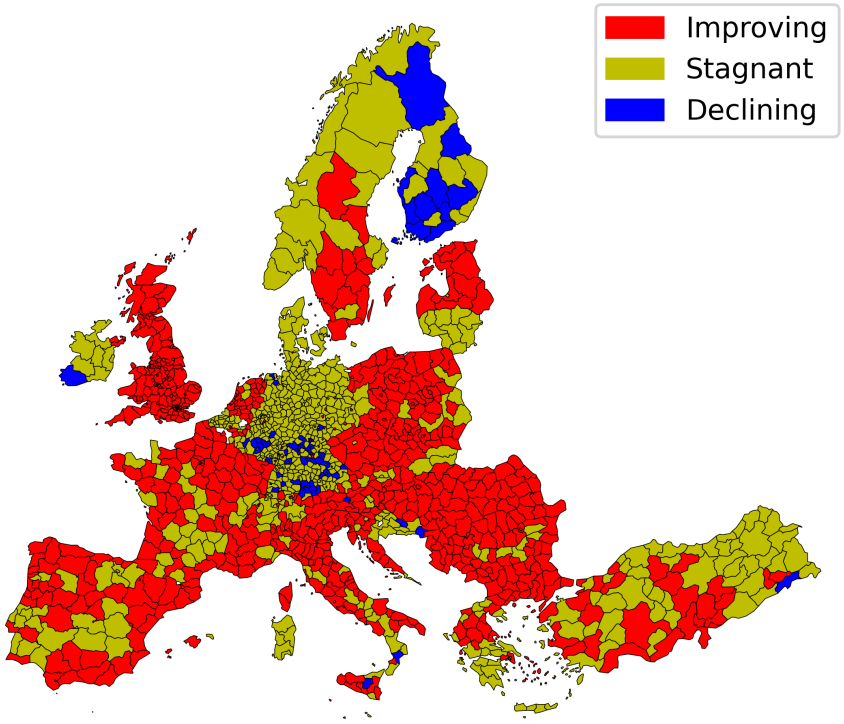

Figure 10: A map colour coded by whether the region is improving (red), stagnant (yellow) or declining (blue).

of possible Ds. This diversity is most evident for the most rural regions, coded NMR-R. It is likely this is genuine diversity as opposed to and artefact of low number statistics, because 294 of the regions are classified as NMR-Rs. In comparison, 264 are classified as NMR-Ss, 329 as NMR-Ms, 381 as MR-Ms, and 198 as MR-Ls.

3.2 Time series analysis

To produce a meaningful time series for a region there must be sufficient data that D can be calculated in least three years. We confirm that this is possible for all the regions within this data set. For each of them the Pearson's correlation coefficient and its p-value is computed. Regions with a significant ($p < 0.05$) positive correlation between time and D are classified as improving regions. Those where there is a significant negative correlation are classified as declining regions. Where there is no significant correlation between time and D the region is considered stagnant. We note that regions with fewer years of recorded data are biased towards the stagnant category, as it is more difficult for low N data sets to meet the p-value threshold. As a result the stagnant classification may be over-represented in the results.

Using this method 768 improving regions, 641 stagnant regions, and 57 declining regions are identified. The geographic distribution of these is shown in Fig. 10, with the improving regions shown in red, the stagnant ones shown in yellow, and the declining regions shown in blue. It is apparent from this figure there is a strong spatial correlation, (confirmed via Moran's I, which is 0.16), with central Europe, the Republic of Ireland, Norway, Finland, Denmark, and eastern Turkey faring the worst. However, on the whole, on individual local scales at least, there has been significant improvement in regional development over the time period of this dataset (2010 - 2021), and comparatively very little backsliding. The high rate of stagnation in many regions presents a concern, however.

The relationship between a region's D and whether it is improving is now examined. The Cumulative Distribution Function (CDF) is calculated for improving, stagnant, and declining regions separately as a function of their average D over the period, and the results are shown in Fig. 11. From this

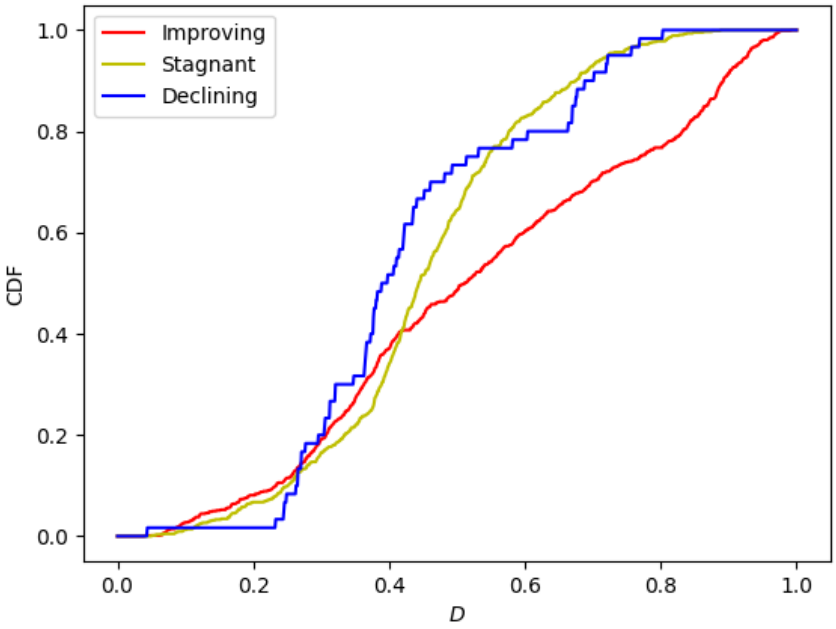

Figure 11: The Cumulative Distribution Function (CDF) of improving (red), stagnant (yellow) and declining (blue) regions compared to their regional development score (D, shown on the x axis).

figure it is apparent that improving regions are disproportionately those that already have high regional development. On the other hand approximately 80 % of both declining and stagnant regions are either vulnerable or underdeveloped ($D < 0.5$).

This poses a concern, as it is an indication of worsening regional inequality, with already well developed regions developing still further on average than their less developed counterparts. However, it is worth noting that while those aforementioned trends are generally evident in Fig. 11 they are not hard and fast rules. While improving regions are *onaverage* those with high D, regions with low D that are improving plainly exist too. Interestingly, there are very, very few declining regions at either extremely low D values (almost none at $D < 0.25$), or extremely high D values (none at $D > 0.8$). Equally, while stagnation is experienced for regions at a very large range of D levels, there are no stagnating (or declining regions) with $D > 0.9$. This means the very most developed regions are all improving, increasing potential concerns about worsening inequality.

The average D across all regions over time is shown in Fig. 12, and the standard deviation of recorded Ds in each year are shown by error bars. There is no obvious correlation between the two variables, and the lack of significance is confirmed with Pearson's correlation coefficient, which finds a p-value of 0.44. This result, contrasted with the previous finding that individual regions are far more likely to be improving ($N = 768$) than declining ($N = 57$) implies a highly complicated and heterogeneous distribution of regional development. It further implies that D is decreasing faster in declining regions than it is increasing in improving regions; this would be necessary given the higher population of improving regions while the average D across all regions is not changing significantly.

The correlation between region's time series trends and their degree of urbanisation is now examined. The fraction of declining regions that are in remote (NMR-R type) is calculated. Likewise, the fraction that are in each of the other four urbanisation classes is computed. These relative fractions are shown in the first column Fig. 13, in which urbanisation class in order of increasing urbanisation is shown on the x axis, and the fraction of regions in each class is shown on the y axis. The fraction of declining NMR-S regions is slightly higher than the NMR-R declining fraction. However, given the relatively low total number of declining regions (57) and the small difference between the declining NMR-R and declining NMR-S fractions this may be due to the stochasticity of low number statistics. There is, however, a convincing peak in the fraction of declining regions that are also of medium urbanisation (NMR-M type).

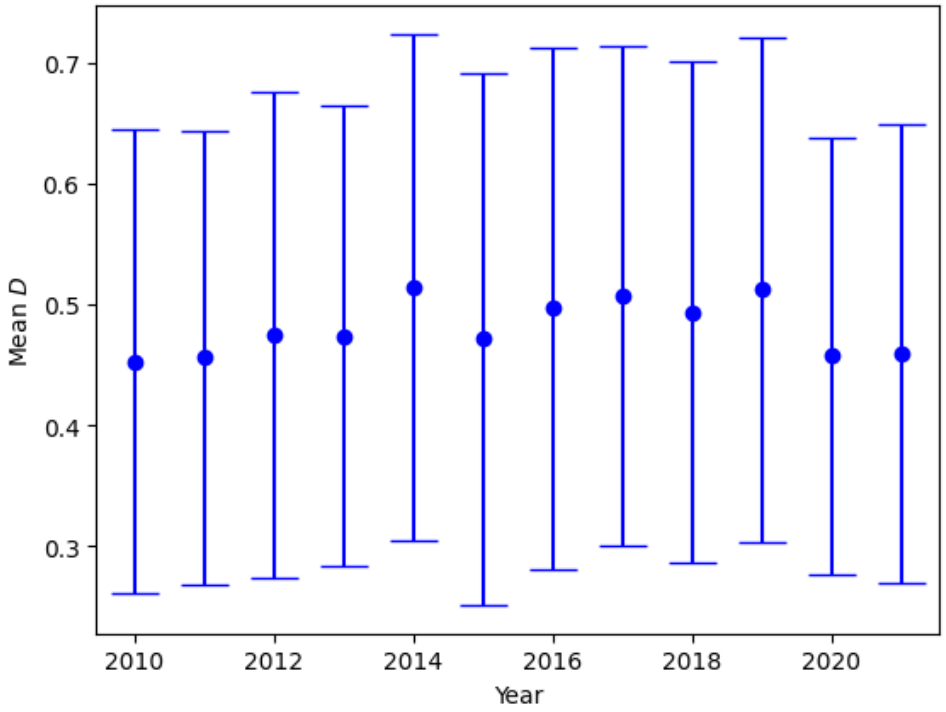

Figure 12: The average D in regions as a function of time. The error bars show the standard deviation of D values within the relevent year.

Likewise, there is also from that point a convincing decrease in declining region fraction with increasing urbanisation; the declining fraction is smaller for MR-R type regions than NMR-M type regions, and still smaller still for MR-L (the most urban) regions. In fact there are almost no declining regions that are also highly urban.

Collectively this data conveys that declining regions are very unlikely to be highly urban (MR-L), somewhat likely to be very remote, small, or part of a metropolitan midsize region, but they are more likely to be in proximity to a midsize/large functional urban area with more than 250,000 people.

The picture for stagnant regions (second column of Fig. 13 is more mixed. There are comparable fractions of stagnant regions in the NMR-R, NMR-M, and MR-M categories, with relative fractions of approximately 0.25 in each of those three classes. The remainder appear fairly evenly split between the NMR-S and MR-L categories. It is not obvious why stagnant regions would appear to be so bimodaly distributed in the NMR-R, NMR-M and MR-L regions compared to the NMR-S and MR-L regions. The lack of a clear correlation between urbanisation and stagnation fraction implies that whatever factors are driving regional stagnation they are either a) not related to how urban regions are, or b) connected to urbanisation in a highly non-linear way.

For improving regions there also does not appear to be a clear bias towards higher or lower levels of urbanisation, though the stark difference between the MR-M and MR-L categories is noteworthy. All in all, these findings collectively demonstrate that the relationship (if there is one) between urbanisation and the rates of improving regions is not straightforward, with no obvious relationship between the two apparent in Fig. 13.

4 Policy implications

Widespread international inequality is observed, with the south and east of Europe faring worse than the north and west. Regional inequality is already a pressing concern for EU policy makers, as evidenced by the EU's regional development fund, and the Horizon Europe funding call which finances this project. This (expected) finding re-emphasises that inequality is persistent and severe, and remains

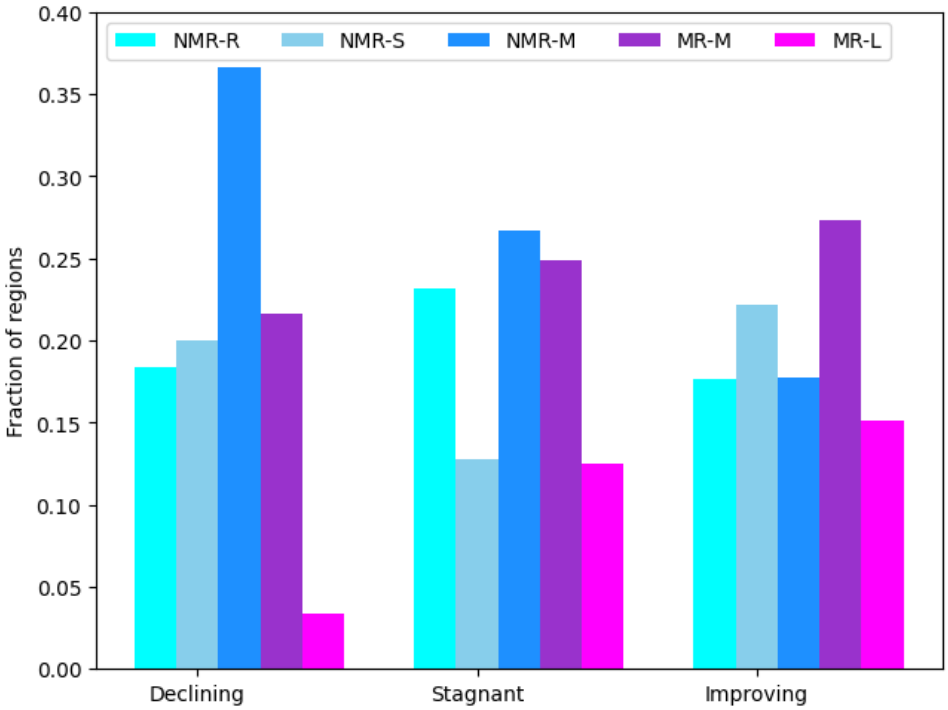

Figure 13: The fraction of declining, stagnant, and improving regions in each of the five urbanisation categories. The region's development trends are shown on the x axis in increasing order, and these are further broken down into degree of urbanisation. The fraction of the regions in each class is shown on the y axis.

worthy of significant political attention and intervention. A high degree of inequality is also observed within countries. National and regional policy makers can use the findings of this paper to identify which regions are most at risk and target interventions accordingly.

Most (52 %) of regions are found to be improving, which is encouraging, however 43 % of regions are stagnant and the remainder are declining, which presents a concern. Further, improving regions are disproportionately those that are already highly developed, whereas vulnerable and underdeveloped make up approximately 80 % of declining and stagnant regions. These results highlight a worrying trend, which left unchecked by systematic intervention will likely lead to strongly increased inequality in the long term.

Stagnant and declining regions are widespread but unevenly distributed. Further investigation is required to evaluate why those specific regions are performing badly, and precisely how that decline is presenting within them. From that, policies can be developed to try to improve their future evolution. It is possible that an examination of the demographics of these regions will shed light on these complex and interrelated questions. For example, we hypothesise that declining regions will have ageing populations. If correct this can be regarded as both a symptom and a cause of regional decline. See section 5 for more information on how such an analysis will be conducted.

Another pertinent and somewhat unexpected finding in this paper is that regional development status (D) is only weakly correlated with urbanisation. This is in contrast to the generally held stereotypes of highly developed cities and crumbling rural outlands (though of course many regions that fit those archetypes also exist). This diversity within the types of regions that are vulnerable or underdeveloped implies that a diverse range of policies will be needed to address each region's particular challenges.

An additional policy area that is thrown into sharp relief by this work is data collection and management policy. There has been considerable progress in this area over the years, however, as outlined in section 2, data collection is still highly inhomogeneous, particularly for social indicators. Increasing the homogeneity of data collection internationally is vital to improve useful international comparisons, the sharing of techniques to identify problems, and the transfer and localisation of policies which prove

to be beneficial.

In the interim, the finding of this paper that development in the economic, social and living environment dimensions are tightly correlated is helpful. It means that in regions where data is sparse, or indicators in one (or more dimensions) are unavailable then development in multiple dimensions can still be estimated and policies directed appropriately.

5 Conclusions and future work

This is a rich data set, which offers an excellent opportunity for gaining insight into the status, and evolution, of regional development across Europe. Further, the subnational (NUTS 3) granularity it contains while also covering the scale of a continent makes it a potentially extremely powerful tool for meaningful comparative analysis across very different current (and historical) political contexts. It also represents an unparalleled opportunity in the PREMIUM_EU project as we seek to understand the mechanisms and choices that drive migration at a subnational level, and as we attempt to model how such migration will evolve in the future.

In this project migration data is being gathered on both international and internal migration from a wide range of sources including national registers, censuses, and social media (specifically Facebook and Instagram). Using this, migration flows broken down by age, sex, country of origin, and education will be modelled at the NUTS 3 level. Using the regional development profiles presented here we will explore how a region's development status in different dimensions impacts the types of migrants that are attracted/repelled by it. For example we may hypothesise that international migrants will be disproportionately attracted to highly economically developed in regions. In contrast we may hypothesise that internal migrants moving away from urban areas may prioritise regions with high living environment and/or social development.

From understanding *what* those migration trends are we can them move on to understanding the *why* they exist, and gain insight into the core motivations of many different types of migrants. These insights can then be combined with the identification of vulnerable regions performed in this paper to support research into effective policies for developing those areas that are most critically in need.

Alongside this work, efforts to improve the Regional Development Characteristics Database will continue. As discussed in section 2.1, more sophisticated methods of disaggregation will be explored. In addition to this, the possibility of using highly complete indicators to model less complete ones will be evaluated. In this evaluation the potential increase in completeness in the dataset will be weighed alongside the robustness of such potential models.

Another avenue of exploration is to re-analyse this regional development data within the context of the Regional Demographics Characteristics Database also produced by this project, and discussed in section 1. This will allow us to answer questions that lie at the crossroads of development, demographics, migration and policy, and thus represents a tremendously exciting opportunity.

References

[1] European Commission. URL: https://ec.europa.eu/regional_policy/funding/erdf_en.

[2] Milenko Fadic et al. "Classifying small (TL3) regions based on metropolitan population, low density and remoteness". In: (2019). DOI: https://doi.org/https://doi.org/10.1787/b902cc00-en. URL: https://www.oecd-ilibrary.org/content/paper/b902cc00-en.

[3] Korrie Melis and Elles Bulder. "Regional development indicators. Creating a database on regional development indicators at Nuts 2 and Nuts 3 levels". In: *The 63rd European Regional Science Association Congress*. July 2024.

[4] OECD. *OECD regional statistics*. URL: Source:%20https://doi.org/10.1787/region-data-en.

[5] Andy Pike, Andres Rodriguez-Pose, and John Tomaney. *Local and Regional Development*. 2nd ed. London, England: Routledge, July 2016.

[6] Jeroen Smits and Iñaki Permanyer. "The Subnational Human Development Database". In: *Scientific Data* 6.1 (Mar. 2019), p. 190038. ISSN: 2052-4463. DOI: 10.1038/sdata.2019.38. URL: https://doi.org/10.1038/sdata.2019.38.