Location and context – analysis of spatial inequalities at different geographical scales. Deliverable 5.3

Melo, Patrícia ; Gaspar, José; Janssen, Heleen; van Ham, Maarten; Andersson, Eva; Malmberg, Bo

Publication date
2019

Document Version
Final published version

Citation (APA)
Melo, P., Gaspar, J., Janssen, H., van Ham, M., Andersson, E., & Malmberg, B. (2019). *Location and context – analysis of spatial inequalities at different geographical scales. Deliverable 5.3*. H2020 project Relocal.

Resituating the Local in Cohesion and Territorial Development

Deliverable 5.3: Location and context - analysis of spatial inequalities at different geographical scales

Authors: Patricia C. Melo (ISEG, Hutton), José Gaspar (ISEG), Heleen J. Janssen (TU Delft), Maarten van Ham (TU Delft), Eva Andersson (UStockholm), Bo Malmberg (UStockholm)

Report Information

Title:	Deliverable 5.3 Location and context - analysis of spatial inequalities at different geographical scales
Authors:	Patricia Melo (ISEG, Hutton), José Gaspar (ISEG), Heleen Janssen (TU Delft), Maarten van Ham (TU Delft), Eva Andersson (UStockholm), Bo Malmberg (UStockholm)
Date of Publication:	March 2019

Project Information

Project Acronym	RELOCAL
Project Full title:	Resituating the Local in Cohesion and Territorial Development
Grant Agreement:	727097
Project Duration:	48 months
Project coordinator:	UEF

Co-funded by the Horizon 2020
Framework Programme of the European Union

Contents

Co-funded by the Horizon 2020
Framework Programme of the European Union

Co-funded by the Horizon 2020
Framework Programme of the European Union

List of Figures

Co-funded by the Horizon 2020
Framework Programme of the European Union

Co-funded by the Horizon 2020
Framework Programme of the European Union

List of Tables

Co-funded by the Horizon 2020
Framework Programme of the European Union

1. Executive Summary

This report provides empirical evidence on the relationship between local area income deprivation and individual socio-economic outcomes using a multi-scale approach. It uses interrelated data on individuals, place and time to investigate the influence of contextual local area income deprivation on individual labour income after controlling for individuals' characteristics and, where possible, family background. To have a better understanding of the consequences of contextual local area income deprivation on individual's outcomes, it is important to consider the suitability of different geographical units both in terms of scale (i.e. from aggregate to very disaggregate) and type of boundary (i.e. administrative fixed boundaries vs more flexible boundaries). Different geographical scales and boundaries may lead to different results, with consequences on the design of public policies and their expected outcomes. Therefore, the main contribution of this work is its ability to define and measure *neighbourhoods* in a more precise or meaningful way to address issues of multiple scales and boundaries by using bespoke neighbourhood measures.

The analyses carried out in the report use geocoded longitudinal microdata for Sweden, the Netherlands and the UK, as well as longitudinal microdata from the EU-SILC for the RELOCAL partner countries with geographical identifiers for NUTS2 regions: Spain, France and Finland. Given the nature of the data available, different types of empirical analyses were developed with varying levels of methodological sophistication and spatial resolution. The "best in class" data, and consequently analyses, were produced for Sweden and the Netherlands, which can be viewed as demonstrators of what can be achieved with access to highly disaggregate geocoded socio-economic information for the whole population. The empirical analyses for Sweden and the Netherlands are particularly interesting because they measure the effect of income deprivation in the local or larger area during adolescence (around aged 15-16) on individuals' labour income as adults (in their late 20s to age 30). The data available for the UK is also of very high quality, but does not allow the same level of spatial detail and is based on survey data (i.e. covers a sample of the population) and thus does not allow linking exposure to local area deprivation in adolescence to adult life income level. The results from the three country-specific analyses on how contextual poverty affects individual labour income have shown that local area income deprivation affects individual labour income level, that is, higher

Co-funded by the Horizon 2020
Framework Programme of the European Union

concentration of poor households is associated with lower individual income. However, this relationship differs depending on the geographical scale at which contextual income deprivation is measured. The effect appears to be most pronounced for lower spatial scales, especially for Sweden and the Netherlands, and to less extent the UK. Scaling up to larger geographical areas, such as NUTS2 regions, the concentrations of low-income households are naturally averaged out, resulting in smaller differences between poverty concentration at these scales. This in turn, reduces the relative importance of the estimated effect of contextual income deprivation on individual outcomes.

As for the EU-SILC analyses, while using longitudinal microdata, they are limited geographically to aggregate NUTS2 regions which hide substantial within-region variation in socio-economic conditions. The main limitation of the EU-SILC dataset is in fact the aggregate nature of the geographies referring to individuals' residential area with data available only at the level of NUTS1 regions for the majority of countries included in the survey. This creates strong limitations to any empirical analyses aiming to disentangle the relative importance of contextual place-specific effects at different spatial scales, particularly lower spatial scales. Nevertheless, it was possible to use the EU-SILC microdata to investigate patterns of income mobility, income inequality and inequality of opportunity across NUTS2 regions and by degree of urbanisation (i.e. large urban areas, small urban areas, and rural areas) of individual's residential location for Spain, France and Finland. Overall, the results indicate that regional differences matter. In particular, the analyses provide some indication of a negative correlation between NUTS2 population size and the degree of upward mobility, in line with the results by degree of urbanisation suggesting less income mobility for large urban areas. In addition, there is considerable variation in income inequality at the regional level, particularly if measured using different income share ratios as opposed to the more general measure of income inequality based on the Gini Index. This means that apparently similar levels of overall income inequality may hide variation in the more local profile of inequality between income shares in the top or bottom sides of the income distribution. Another interesting result emerging from the EU-SILC analyses is that there are also regional differences in the degree of inequality of opportunity, that is, relative importance of individuals' *circumstances* (i.e. family background, gender), as opposed to individuals' *effort* (i.e. factors individuals can influence), to individual economic outcomes. The degree of inequality of opportunity was found to be larger for urban areas compared to small urban areas and rural areas, as well as considerable differences across NUTS2 regions.

The work carried out in this report also has some limitations. It measures residential poverty and deprivation in terms of low-income concentration, but it is important to note that income is only one dimension of poverty and while it would have been preferable to adopt a more multiple dimension definition this was not possible for data reasons. Furthermore, while the multi-scale approach shows that inequality is a multi-scale problem, on its own it cannot explain which mechanisms operate at different levels; achieving this requires combining them with detailed case study analysis.

One of the main conclusions from this report is that in order to have a better understanding of residential context, in particular area income deprivation, on individual socio-economic outcomes, it is important to measure and test the relationship at different geographical scales. However, the approach implemented in this report for Sweden, the Netherlands, and the UK can only be applied when geocoded data are available for very small spatial units and such data are still unavailable in many countries. Consequently, one very important conclusion and recommendation from the work carried out in this report is the need to improve the availability and access to socio-economic geocoded data at very low scale for more countries. Without this type of information, it is not possible to provide guidance to policy makers on the more appropriate scales for public policy intervention.

Co-funded by the Horizon 2020
Framework Programme of the European Union

2. Introduction

2.1. Background and motivation

This report forms Deliverable 5.3 of the EU Horizon 2020 research project ‘Resituating the local in cohesion and territorial development’ - RELOCAL. Work Package 5 addresses the spatial inequalities at multiple geographical scales, using methods that do not depend on administrative regions. The latter is important because the efficiency of specific policy interventions directed towards spatial inequalities is scale-dependent and hence it should be based on well-defined and meaningful measures of spatial variation in living conditions. The work carried out in Task 5.3, and which is documented in this report constituting deliverable D5.3, is interested in providing answers to the question of how *place* impacts on the socio-economic chances of individuals. The term *place* refers to the geographies where individuals live, which can vary by nature (e.g. administrative, statistical, functional, etc.) and geographical scale (i.e. from very disaggregate residential blocks or neighbourhoods to large regions).

Individual inequalities in the level of socio-economic achievement (e.g. employment, income) result both from differences in the characteristics of individuals and differences in the places where people live/work. It is important to start the report by clarifying our use of the terms *area or place effects* and *neighbourhood effects* as we will use them interchangeably in this report although their specific meaning can vary according to discipline. By these terms we refer to the influence of residential location on individual outcomes, and in particular income in the case of our work. While the term area or place implies a spatial or geographic environment or location, the term neighbourhood is often used to mean the belonging to a given group sharing values, behaviours or outcomes, and we know in the case of neighbourhood effects that these groups also share a geographical location, that is, there is correspondence between the social and spatial dimensions of the group (i.e. between neighbours both in the physical and social meanings of the term). We use the terms *area or place effects* and *neighbourhood effects* in this sense, that is, of the correspondence between the social, economic and spatial dimensions

Co-funded by the Horizon 2020
Framework Programme of the European Union

To reduce inequalities it is therefore important for public policies to be informed about the relative importance of individual (i.e. 'people') effects and contextual (i.e. 'place') effects. There are complex selection mechanisms influencing people's decisions about where to live. These mechanisms can operate at different spatial scales, ranging from local labour markets where worker-firm matching occurs, to more localised social networks within the residential environment or neighbourhood of individuals. To have a better understanding of the consequences of spatial inequality on individual's outcomes, it is therefore important to consider the suitability of different geographical units both in terms of scale (i.e. from aggregate to very disaggregate) and type of boundary (i.e. administrative fixed boundaries vs more flexible and meaningful boundaries). Different geographical scales and boundaries may lead to different results, with consequences on the design of public policies and their expected outcomes. For example, poverty can be concentrated in particular regions, cities, or neighbourhoods. Spatial inequalities within regions might be much larger than between regions, which is important for the development and implementation of policy measures to counter inequality. Analysing spatial inequality and poverty concentration at an aggregate geographical scale may hide considerable variation at a smaller geographical scale. Furthermore, and as noted above, besides the issue of scale, administrative boundaries may not necessarily correspond well to the reality of income inequality and poverty incidence.

2.2. Current state-of-the-art of the empirical literature

Existing research on area effects has considered a wide range of outcomes, including education, employment status, occupation, income, health, etc. For a review of the literature see Ellen and Turner (1997), Galster (2002), Dietz (2002), Durlauf (2004), van Ham and Manley (2010). There appears to be conflicting views on the presence and importance of area effects between disciplines and methodological approaches, with qualitative studies using field interviews tending to find evidence of neighbourhood effects, while econometric studies based on observational data tend to find mixed evidence and those based on quasi-experiments generally find little or no evidence in favour of neighbourhood effects (e.g. Durlauf, 2004, Bolster et al., 2007). Given the wide differences in methodologies, type of data, and outcomes studied, it is difficult to make conclusive comparisons, but the prevailing view seems to be that in the absence of experimental or quasi-experimental studies (either for

Co-funded by the Horizon 2020
Framework Programme of the European Union

ethical or practical cost reasons), quantitative studies using longitudinal geocoded microdata offer better chances of overcoming identification issues relating to self-selection bias, unobserved heterogeneity, and reverse causation, which prevent any conclusions about causal effects (e.g. Durlauf, 2004, Cheshire, 2007).

Therefore, and despite the abundant and growing body of research, important challenges persist and need addressing in order to move the literature, and its contributions to policy and practice, forward. van Ham and Manley (2012) discuss ten challenges for neighbourhood effects research, some of which are directly addressed by the work carried out in Task 5.3, namely the identification and measurement of neighbourhoods in a more meaningful way to address issues of multiple scales and boundaries and the use of bespoke data to investigate neighbourhood effects.

There has been progress addressing some of the main estimation issues, namely those arising from residential sorting and the choice of relevant geographical units. The methods considered in previous studies can be implemented either separately or in combination, and typically include one or more of the following approaches: sample restriction, longitudinal data and individual fixed-effects, instrumental variables, use of a control function based on hedonic house prices, and explicit modelling of neighbourhood choice (see Appendix A).

2.3. Objectives

To provide a better understanding of the importance of place-related contextual effects, measured at different spatial scales, on individuals' economic outcomes, Task 5.3 uses interrelated data on people, place and time to investigate the influence of contextual geographical characteristics on individual economic outcomes after controlling for individuals' characteristics. Data requirements for Task 5.3 were clearly identified in Task 5.1 whose main aim was to assess the availability of geocoded longitudinal individual level data with respect to social and spatial inequality. Essentially, the key data requirements were: geocoded data, where the spatial scale of the geographical units may range from very low to very high (we needed low spatial scale for Task 5.2 and 5.3); longitudinal data, i.e. the information is collected for the same subjects over time/at multiple times; and microdata, i.e. data at the level of individual persons or households.

The empirical work carried out in Task 5.3 focuses on individual income as the main outcome of interest. We develop empirical models based on individual longitudinal data to measure how much of the differences in individual's income levels can be related to contextual area degree of low-income or poverty concentration. We acknowledge that income is only one of the dimensions of poverty and while we would prefer to use a more multiple dimension definition this was not possible for data reasons.

In summary, task 5.3 has two main objectives: i) to investigate the relative importance of area-level income deprivation on individual labour income, and ii) whether measuring contextual area income deprivation at different spatial levels and/or using different types of geographies affects the results.

2.4. Spatial coverage

The analyses carried out as part of Task 5.3 use national geocoded longitudinal microdata for Sweden, the Netherlands and the UK. Although similar geocoded longitudinal microdata are also available for Finland, data access limitations did not allow WP5 researchers to carry out the analysis for this country. As discussed in Task 5.1, we also considered the suitability of using pan-European longitudinal microdata, in particular the EU-SILC. Although the EU-SILC fulfils several of the necessary data requirements mentioned above, the main limitation is the very aggregate nature of the geographies referring to individuals' residential area. For the majority of countries data are only available for NUTS1 regions, with only a few (Spain, France, Finland, Check republic) having data for NUTS2 regions. We therefore had initially decided to discard EU-SILC for the purpose of Task 5.3. However, and given the richness of the questionnaires, we reconsidered the initial decision and have also carried out analyses for the three RELOCAL countries with NUTS2 data available in EU-SILC. For the reasons stated earlier, the analyses carried out for these countries are not directly comparable to those performed using the country-specific databases, which in the case of Sweden and the Netherlands consist of register data whilst in the case of the UK consist of survey data. Furthermore, as discussed later in the empirical methods section, other EU-SILC specificities have influenced the types of analyses carried out with it.

Co-funded by the Horizon 2020
Framework Programme of the European Union

2.5. Main contributions

In the work carried out in Task 5.3, we define and measure neighbourhoods in a more meaningful way to address issues of multiple scales and boundaries by using bespoke neighbourhood measures. There is no consensus on what the appropriate boundaries and scale of geographies should be, the current understanding is that administrative spatial units are generally not fit-for-purpose and that area effects may operate at different scales which are likely to vary according to the relevant mechanism being studied, and can range from immediate local neighbourhoods to the local labour or housing markets and wider regional economies. Some studies have used bespoke measures of neighbourhoods drawn around household's homes using nearest neighbour thresholds (e.g. Buck, 2001, Bolster et al., 2007, Hedman et al., 2015). These measures allow us to test the question about whether the socioeconomic status of one's neighbours (e.g. having poor or rich neighbours) impacts on one's own socioeconomic outcome (in the case of WP5 this refers to individual income). However, this approach can only be applied when geocoded data are available for very small spatial units and such data are still unavailable in many countries. The work in Task 5.3 will contribute to this challenge by using different types of boundaries in the estimation of local area income deprivation effects. More specifically, it will use flexible boundaries obtained through the bespoke neighbourhood approach implemented in Task 5.2., as well as more conventional boundaries based on administrative geographies.

2.6. Structure of the report

This report is structured as follows. Section 3 describes the data and empirical methods used in Task 5.3 following on from the outcomes from the previous tasks 5.1 and 5.2. Sections 4 to 7 present and discuss the empirical analyses carried out using the pan-European data from EU-SILC (section 4), the United Kingdom (section 5), Sweden (section 6), and the Netherlands (section 7). Finally, section 8 provides the main conclusions.

3. Overview of empirical approach: data and methods

3.1. Introduction

The core of the work in Task 5.3 consists of estimating longitudinal regression models using the secondary data sources identified as being fit-for-purpose in Task 5.1 (see deliverable D5.1) as well as the bespoke neighbourhood measures of income inequality developed in Task 5.2 (see deliverable D5.2).

In this section we describe the empirical methods implemented based on the pan-European and national-level datasets constructed from the EU-SILC and the three national-level datasets for Sweden, Netherlands and the UK. We use multiple methods depending on the data source, including the analysis of the distribution of individual income within-, and between-regions, the construction of indicators of income mobility and inequality, the construction of indicators of regional inequality of opportunity, the development of longitudinal individual-level regression models that measure the relative importance of people and place effects on individual income level. The later of the methods, i.e. longitudinal microdata regression models, allows us to (attempt to) control for some of the main identification issues faced by researchers, as discussed in the previous section of the report. The following section provides a brief overview of the main empirical methods used.

3.2. EU-SILC longitudinal microdata analyses

The EU Statistics on Income and Living Conditions (EU-SILC) includes data for a wide set of variables collected by Member States in their respective national surveys. It contains both cross-sectional and longitudinal samples, where the latter follows a 4-wave rotational design. In order to access the EU-SILC microdata the RELOCAL team at ISEG submitted a scientific research proposal and accepted all the required terms of use and individual confidentiality declarations. For the purpose of Work Package 5, we have used the panel data component of the EU-SILC for the longest period of data available to date (2005-2016). Given the focus of Task 5.3 on area-level contextual effects, the objective was to investigate geographical heterogeneity across NUTS2 regions and by degree of urbanisation. As a result, we considered only the RELOCAL countries in the EU-SILC microdata files for which NUTS2 level information is available: Finland, France, and Spain.

EU-SILC longitudinal data are reported in four different files: (i) a household register (D) file, that contains basic information from households regarding the selected sample, such as country, region (NUTS1 or NUTS2), and degree of urbanisation,[1] among others; (ii) a household data (H) file that contains more specific data (household income, social indicators such as social exclusion, housing, among others); (iii) a personal register (R) file that consists of basic data on individuals such as country, year of survey, sex, among others; and (iv) a personal data (P) file containing more specific information (labour market, health, income, among others) for all household members aged 16+ for whom the information could be completed (in other words, the individuals in P-fie are a subset of those in the R-file. For each country and each year, the D- and H-files have unique household identifiers, whereas the R- and P-files have unique identifiers for individuals. Additionally, the R-file also contains household identifiers matching the ones in the D- and H-file so that both household and individual level data can be linked together, and form one master data set obtained from merging all the aforementioned files.

Appendix B provides a detailed description of the EU-SILC data files and the various data cleaning and management procedures undertaken in order to obtain a master dataset suitable for the empirical analyses. The appendix also describes the data management operations carried out for the different countries included in the empirical analyses.

The outcome variable of interest in our analyses is labour income, more specifically, employees' earnings from work. As described in Chap. 24 of Atkinson et al. (Atkinson et al., 2017), labour income in the EU-SILC data is simply the annual gross (net) employee cash or near income: "the monetary component of the compensation in cash payable by an employer to an employee, and it includes the value of any social contributions and income taxes payable

[1] There are three degrees of urbanisation: (i) Large urban areas - contiguous grid cells of at least 1 500 inhabitants per squared km and at minimum population of 50 000; Small urban areas - clusters of contiguous grid cells of 1 squared km with a density of at least 1500 inhabitants per squared km and a minimum population of 5000; and (iii) Rural areas - grid cells outside urban clusters.

Co-funded by the Horizon 2020
Framework Programme of the European Union

by an employee or by the employer on behalf of the employee to social insurance schemes or tax authorities. As a result, we always refer to labour income when using the term income in the report. The variables regarding individual income (contained originally in the P-file) report to “Employee cash near-cash income” as follows: PY010G - Total remuneration in cash or in kind by an employer to an employee in return for work done by the latter during the income reference period, before deduction at source of any taxes or social contributions.[2] PY010N - The net income corresponds to the gross component but without any deductions at source such as taxes or social contributions.

In order to carry out analyses over time and across countries, we merge the EU-SILC income data with the latest information on Harmonized Indices of Consumers Prices (HICP) provided by Eurostat[3] to deflate gross income for cross-year comparison and convert all income values to constant prices of 2015. The reference sample for the empirical analyses of labour income consists of working age people, aged between 16 and 65 years old.[4] However, we compute the variable age by subtracting the year or birth (PB140) to the year of the survey and drop all observations such that age>65. We also look at the variable RB170 “Main activity status during the income reference period” and drop all observations referring to non-working individuals during the income reference period. We further remove all individuals with gross income equal to zero or missing values during the income reference period. To account for the presence of outliers in the distribution of labour income, we use data on minimum wages from Eurostat[5], which is calculated based on 12 monthly payments per year. We deflate these minimum wages to constant prices of 2015 and remove all observations with gross labour income (i.e. *eginc*) lower than 3/5 of the annual minimum wage. This rule of thumb allows us

[2] For France, gross labour income is actually part collected net, part collected gross, i.e., it is net of tax on social contributions.

[3] See https://ec.europa.eu/eurostat/web/hicp/data/database.

[4] According to (Mack, 2016) there was a variable in EU-SILC RX010 dubbed “age at the date of survey”; it has been removed from EU-SILC data.

[5] See https://ec.europa.eu/eurostat/statistics-explained/index.php/Minimum_wage_statistics.

to account for the existence of part-time workers and thus leads to a lower loss in the number of observations. Since there is no minimum wage in Finland, we exclude observations with reported income lower than half of the average wage, which lies in accordance with Statistics Finland (Eurofound, 2009). In addition, following Alperin et al. (2013), we remove the upper extreme values of the income distribution by dropping values that are 25% higher than the 99th percentile, for each year of the sample between 2005 and 2016. Finally, the cleaned EU-SILC datasets for the three countries were then merged with NUTS2-level contextual data obtained from Work Package 2. The analysis of the EU-SILC country-specific datasets comprises four separate empirical investigations reported and discussed in detail in Appendix D, and summarised in chapter 4.

3.3. Country-specific analysis for the United Kingdom

- *Longitudinal microdata*

We use microdata data from the UK Household Longitudinal Study (UKHLS), which is a large multipurpose annual longitudinal survey that collects data for individuals and households. At the time of this study, data were available for waves 1 to 6 covering the period from 2009 to 2015.[6] The survey contains a series of modules (some of which are applied on a rotating basis) including a wide range of topics referring to individual and household demographics, socio-economics, health and well-being, personal transport, consumption and housing expenditure, and environmental attitudes and behaviours, among other topics. For the purpose of this work, we are interested in the data relating to the demographic and socio-economic characteristics of individuals and their households, including information on parental socio-economics which can be used to account for potential inter-generational transmission of (dis)advantage, as well as information on individuals' residential location in order to link in contextual information about it at different spatial scales (see section 3.2.4.2 below). The

[6] We use the dataset SN6931 (University of Essex. Institute for Social and Economic Research, 2016).

empirical analysis carried out uses a sample of working-age individuals in full-time employment, thus excluding students, retirees, and the unemployed from the analysis. This is the relevant group of individuals for the outcome of interest in the study, that is, work-related labour income.

The UKHLS provides a rich set of data for individual and household demographic and socio-economic attributes, including age, gender, marital status, highest qualification attained, employment status and regime, income, occupational and industrial affiliation, number and age of children in the household, etc. There are also some variables about respondents' family background, including whether the father/mother was at work when the individual was aged 14 years old, father's/mother's occupation when the individual was aged 14 years old, and father's/mother's education. There are also questions about the national and country of residence of the parents, parents' ethnic group, and whether the individual was living with his/her parents when aged 14 years old. Including these variables in the empirical regression models would help control for inter-generational transmission of (dis)advantage. Unfortunately, these variables are only asked to a sub-sample of the respondents in wave 1 and remaining original sample members in at wave 2. This means that there are many missing values for these variables, which creates a strong limitation to their inclusion in the specification of the individual labour-income regression models.

As discussed earlier, the focus of the empirical analyses carried in task 5.3 of work package 5 is on the role of geographical context, specifically in terms of concentration of low income, on individual income outcomes. Furthermore, we want to know whether this relationship differs depending on the type of geography and spatial scale we use. In order to test this hypothesis, we measure geographical concentration of low income using administrative geographies (local authorities, LAD) and census-based small areas, as well as the bespoke geographies constructed in task 5.2 (and reported in Deliverable 5.2). These measures were linked to the UKHLS dataset, and are described with more details in the following section.

- *Data for local area income deprivation*

The measures of local area income deprivation for the UK refer to England and Scotland, the two countries studied in Task 5.2 and which have case studies in the RELOCAL project. For details on the computation of the bespoke measures of neighbourhood income deprivation and the geographies underlying the analysis, we refer the reader to Appendix E. The bespoke geographies for income deprivation computed in Task 5.2 using the EquiPop software considered multiple spatial scales, staring from the proportion of people with income below 60% of the median income among the nearest 200 people (400, 800, 1600, 3200, …), up to the proportion of people with low income among the nearest 204,800 people. By increasing the scale, the contextual variable of each grid cell (which is a proxy for a residential location) measures poverty for a larger population, and by definition also a larger geography. For larger k-neighbour thresholds, these bespoke geographies can reach sizes similar to those of administrative geographies (e.g. local authorities or council areas). Furthermore, in the specific case of England and Scotland, and compared to Sweden and the Netherlands, the underlying geographies for the income data start with relatively large building blocks (e.g. MSOAs for England), which already contain more households than some of the EquiPop thresholds (e.g. k=200, …, 800) leading to the exclusion of these lower spatial scales. Consequently, the econometric analysis carried out for England and Scotland considers measures of income poverty for the following geographies:

- Measures of income poverty for the bespoke geographies computed for different spatial scales using EquiPop's k-nearest neighbours approach, as per Task 5.2;
- Fixed boundary official small area geographies, DZs in the case of Scotland and MSOAs in the case of England; and
- Fixed boundary and more aggregate administrative geographies based on local authorities or council areas (i.e. municipalities) for both countries.

- *Modelling strategy*

Overall the estimation strategy combines the use of alternative estimators and samples. In particular, we implement standard pooled OLS and panel data type estimators such as individual-fixed effects, random-effects and the correlated random-effects (i.e. the Mundlak's correction of the random-effects model, (Mundlak, 1978)). When data are available we also include controls in the model specification for the level of satisfaction with residential location and individual's intention to move home. These are all self-reported variables and thus reflect individuals' subjective perceptions and appreciations about their residential area. Given the data limitations enunciated earlier, it is not possible to experiment with more sophisticated estimators such as instrumental variables (IV) or the use of propensity scores.

3.4. Country-specific analysis for Sweden

- *Longitudinal microdata*

In order to analyse local area income deprivation on individuals' labour-income later in life we have used register data in an analyses of Sweden. The focus of this study is if income deprivation in the local or larger area during adolescence influence an individuals' future in the form of earned income in their late 20s. All data in the Swedish country-specific longitudinal microdata analysis originates from Statistics Sweden's registers in a project called Geographical context covering the years 1990 to 2016. Data is accesses through an on-line system called MONA (Geostar, 2015).

- *Data for local area income deprivation*

We study the 1986 cohort. Residential context is measured in 2001 when the cohort is 15 years of age. Since the *exposure* time in youth is important for an assessment of later effects on outcomes we selected those that had geo-coordinates in the Statistics Sweden registers for the years 1999, 2000 and 2001. The surrounding context was based on a population 25 years and older, and its share having an equalised disposable income less than 60 percent of the median (for procedure and variable please see Nielsen et al., 2017). This corresponds to the EU at risk of poverty measure. Further, the context of poverty was computed at different scale levels as the share of at risk of poverty individuals among the closest 200, 1600, 12800, 51200 and 204800 persons (to compare measures see D5.2 report by Janssen and van Ham, 2018). That is, the local geographical area including 200 neighbours ranging to a large city or region comprising over 200 thousand inhabitants. In the results section (see section 5) this measure is referred to by using the concept of 'neighbourhood' and 'neighbours', though the not so technical use of the word usually denotes both larger and smaller in population sizes. Yet another test of context effects is made using administratively delimited NUTS2 areas (8 regions) and NUTS3 areas (21 regions equivalent to Swedish counties). Thus, contextual effects of the adolescent residential context of youths on early adult life outcomes are tested for altogether seven geo-levels.

Co-funded by the Horizon 2020
Framework Programme of the European Union

- *Modelling strategy*

In the regression models we control for a number of individual and family characteristics: sex, parental tertiary education, family type (single mother families), parental employment, family social assistance, parent per person disposable income (1000s of Euro), non-European background (one or two parents born outside Europe), see Nielsen et al. (2017). For the data procedure of locating every individual's closest number of neighbours at different scales we used a script called *geocontext* (Hennerdal, 2019). The total sample in the cohort is a little less than 104,000 individuals born in 1986 and the context is every person residing in Sweden in 2001 over the age of 25. The regressions were run in the software STATA with a script kindly shared by Heleen Janssen, TU Delft.

3.5. Country-specific analysis for the Netherlands

- *Longitudinal microdata*

For the Netherlands the data source is the Social Statistical Database (SSD, or Social Statistisch Bestand [SSB], see (Bakker, 2002, Houbiers, 2004). The SSD data covers the entire population of the Netherlands, from 1999 – to 2017 and contains data from a range of government registers. The SSD consists of a number of linked registers including demographic, socio-cultural, and socio-economic characteristics of the population. Although the name suggests it is one dataset, the SSD consists of several datasets which can be linked. For each individual basic information is available, such as gender, age and country of birth, but also information on life events such as marriage, divorce and child birth. The data can be enriched with information from other registers including employment status, income, school results of children, and for example criminal convictions. It is also possible to link register data to survey data. The data is geo-coded at the level of 100 by 100 meter grids for the whole country, which can be easily aggregated to larger geographical scales. Administrative geographies are also available in the data. The data can be accessed through a secure remote access facility which has been set up by Statistics Netherlands. Under strict conditions researchers may be granted access to the microdata (see https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research/requirements-for-remote-access). The

Co-funded by the Horizon 2020
Framework Programme of the European Union

data from the Netherlands, and the conditions of access, are very comparable to the Swedish register data.

Individual and parental demographic and socio-economic characteristics: We used two individual outcome variables: individual earned income at age 30 and obtained educational level. *Individual earned income at age 30* is measured as individual income from work. In order to facilitate comparison of the results between the Netherlands and Sweden, we calculated percentiles. These range from 1 to 100, indicating to what income percentile the individual belongs, and therefore the relative income position in the 1987 cohort. However, as income can fluctuate, especially around age 30 due to having children, we included income from age 25 to 30. We calculated percentiles for every year, and then for every individual we took the highest income percentile between age 25 and 30.

Educational level is measured in years. The Netherlands has a highly stratified educational system in which the choices of a field of study are made as early as age 12. Children attend primary school from the age of 4 to 12. In their final year, based on a national test and the teacher's recommendation, they are advised which type of secondary education they should pursue. There are three types of secondary education. One option is lower vocational training (4 years), which gives access to intermediate vocational training (1 years) at the upper secondary level. Two other options are secondary general education (5 years) and pre-university education (6 years). Only the pre-university track gives direct access to university (4-6 years). Al three tracks give access to universities of applied sciences (4 years). In order to make the Netherlands data comparable to the Swedish data, we converted the obtained educational level to years if education.

As individual level predictors of income we included sex (with female as reference category), and a non-European migration background, which indicates whether at least one parent was born outside of Europe. We included a set of family and parental characteristics as predictors of individual income. We included a dummy variable that indicated whether the individual at age 16 was living with their single mother. Another dummy variable indicated whether the family received social allowance when the individual was 16. Household income in thousand Euros was included as a continuous variable. Parental tertiary education was included as a

Co-funded by the Horizon 2020
Framework Programme of the European Union

categorical variable with 3 categories indicating whether no, one or both parent(s) had tertiary education. Parental unemployment was also included as a categorical variable with 3 categories indicating whether no, one or both parents were unemployed when the individual was 16 years old.

- *Data for local area income deprivation*

The lowest geography available in the Dutch register data is 100 by 100 meter grid cells. This data can be made available to researchers under strict conditions by Statistics Netherlands. The individual level longitudinal register data is geocoded so that for each individual in the data it is known in which grid cell he or she lives. As this data covers the whole population of the Netherlands, it is possible to construct contextual characteristics on the level of 100 by 100 meters and higher. Using the 100 by 100 meter grid cells as building blocks, it is possible to aggregate the data to higher spatial scales. In addition to grid cells, administrative geographies are also available in the data, including neighbourhoods, postal code areas, municipalities and NUTS units. For this report, we measure local income deprivation at multiple spatial scales, ranging from very local level to regional level.

Using EquiPop, a specialized software-program for the calculation of the k-nearest neighbours, we constructed individualized egocentric neighbourhoods. The software has been developed by John Östh at Upsala Univeristy (http://equipop.kultgeog.uu.se). The k-nearest neighbour approach, as used in the EquiPop software, provides a tool to draw neighbourhoods at different geographical scales for different types of detailed geographical data. The computation of measures of spatial inequality are based on individualised scalable neighbourhoods, based on fixed population counts. For the current report, we used different scales, ranging from the 200 to the 51,200 nearest neighbours. Preferably, the building blocks which are used as a starting point for the EquiPop analyses are very small and regular. Ideally small grids, such as 100 by 100 meter grid cells or equivalent are used. Individual level data from government registers, or from census data then needs to be aggregated to these small spatial units. The regulations for use of EquiPop software includes the prohibition to profit from the use of the software, for instance users may not sell research reports, presentations and other forms of output and analyses that were produced via EquiPop

Co-funded by the Horizon 2020
Framework Programme of the European Union

(http://equipop.kultgeog.uu.se/Legal/untitled.html). In principle the software is open access for academic research purposes and student work. For details on the computation of the bespoke measures of neighbourhood income deprivation we refer the reader to D5.2 Report on multi-scalar patterns of inequalities (Janssen and van Ham 2018).

As we were particularly interested in the relationship between area deprivation at multiple geographical scales and individual income, for the current report we estimated the effects of contextual poverty at scales ranging from the 200 to the 204,800 nearest neighbours. The building blocks which are used as a starting point for the EquiPop analyses are 100 by 100 meter grid cells. In addition to the bespoke measures we also used administrative NUTS3 and NUTS2 units.

For the Netherlands, we used an indicator of poverty based on the Eurostat definition of the at-risk-of-poverty rate, which is defined as the share of people with an equalised disposable household income below 60% of the median income. The individualized disposable income is obtained from the Netherlands Social Statistical Database (SSD) from Statistics Netherlands. For each geographical scale, using EquiPop, the ratio of individuals of 25 years and older with a low income was calculated.

- *Modelling strategy*

As we were interested in the relationship between neighbourhood deprivation in adolescence and income at age 30, we used the 1987 birth cohort. These individuals were 30 years old in 2017, the last year for which register income data was available at the time of writing this report. We included neighbourhood characteristics from the year 2003, when these individuals were 16 years old.

We estimated the relationship between contextual poverty at age 16 and income at age 30, while controlling for a range of individual and family characteristics. Parental socio-economic characteristics are important to include in this model as they are both related to the type of neighbourhood where the family lived when the individual was 16 years old and to the socio-economic outcomes of the child later in life. The total sample is 158,561 individuals born in 1987 and the data used to construct the contextual measures comprised of every individual of 25 years and older living in the Netherlands in 2003.

4. Results from the EU-SILC microdata analyses

As described in section 3.2 above, we carried out four separate analyses based on the EU-SILC microdata for Spain, France, and Finland. In this section, we present only the main findings, while the full set of results is provided in Appendix D.[7] The names of the NUTS2 regions and a map showing their location within each country are provided in Appendix C.

4.1. Regional variation in income mobility patterns

We analysed the degree of income mobility across NUTS2 regions and by degree of urbanisation (i.e. large urban areas, small urban areas, and rural areas). Measuring income mobility over time is important as it can help to evaluate the extent to which there is upward/downward social mobility along the income distribution. Given the four-wave rotational design of EU-SILC, we can only study individuals' income trajectories up to a maximum of four years (if individuals respond to the survey every year). We consider income mobility for 2-year (i.e. transitions between t-1 and t) and 4-year (i.e. transitions between t-3 and t) income trajectories.

We conclude from Figure 11, Figure 12, and Figure 13 for Spain, France and Finland respectively (see Appendix D), that the level of income mobility over the 2-year and 4-year periods differs according to the degree of urbanisation of the residential area and tends to be higher for less densely populated areas. For example, in the case of Spain, the percentage of workers that moved up two or more deciles in rural areas is 15% (21%) compared to 12% (19%) in large urban areas between t-1 and t (t-3 and t). This is an interesting result given the common perception that cities, in particular, large cities offer great opportunities to "climb the socioeconomic ladder".

[7] More detailed results are available as supplementary material upon request for the interested reader. These include all transition matrices for both gross and net income mobility between t-1 and t and between t-3 and t, at the national level, at the regional level, and by degree of urbanisation. It also contains all the tables regarding income inequality measures both at the national and disaggregated levels. These include the Gini index and the following income share ratios: P90/P10; P90/P50; and P50/P10.

Co-funded by the Horizon 2020
Framework Programme of the European Union

When considering the results for regional-level patterns of income mobility we can conclude that there is considerable regional variation in the three countries, and that there appears to be greater variation in the level of mobility across regions in France, followed by Spain, and Finland. However, discrepancies in upward mobility (1 or more deciles) across regions were very high in Spain and France and lowest in Finland. Furthermore, there is some indication of a negative correlation between regions' population size and the degree of upward mobility in all countries, which corroborates the previous findings regarding mobility by degree of urbanisation.

4.2. Regional variation in income inequality patterns

We now turn to the results of for the level of income inequality across NUTS2 regions and by degree of urbanization based on the analysis of different indicators such as the Gini Index (GI) and income share ratios P90/P10, P90/P50, and P50/P10.

The analysis of income inequality, based on the Gini Index, by degree of urbanization of individual's residential area indicates that the main trend for the overall period from 2005 to 2016 is that income inequality tends to be higher for large urban areas and lower for rural areas, with intermediate values for small urban areas. All countries experienced a decrease in inequality between 2015 and 2016, and inequality is higher in more densely populated areas throughout the period in France and Finland, although the discrepancies are more pronounced in the former. In Spain, between 2010 and 2011 and after 2014 inequality was higher in small urban areas compared to large urban areas.

We also observe considerable variation in income inequality at the regional level, particularly the income share ratios, much more than the gini index, which indicates apparently similar levels of overall income inequality may hide variation in the more local profile of inequality between income shares in the top or bottom sides of the income distribution.

We also considered the association between degree of income mobility and income inequality across regions following Shorrocks' hypothesis (Shorrocks, 1978) that higher mobility is associated with lower income inequality. The evidence gathered from simple scatter plots and pairwise correlations between measures on income inequality and degree of income mobility, in particular upward mobility, give only limited support the Shorrocks' hypothesis. Future

Co-funded by the Horizon 2020
Framework Programme of the European Union

analysis could explore and test this relationship in greater detail (it was not the scope of Task 5.3).

4.3. Regional variation in inequality of opportunity

In this section we consider the extent of inequality of opportunity (IOp) at the national level, across NUTS2 regions, and by degree of urbanisation by estimating regression models that allow deriving measures of ex-ante inequality of opportunity. By inequality of opportunity we mean the difference in individual economic outcomes that result from individuals' *circumstances* (i.e. family background, gender) assuming similar levels of *effort* (i.e. factors individuals can influence). This analysis is based on Stata's user-written command iop Juárez & Soloaga (2014) for Spain and France. The IOp analysis produces a relative measure of the income inequality resulting from individual circumstances, and decomposes it into the different elements of one's circumstances giving an indication of their relative importance.

The results for Spain show that at the national level 19% of the variation in labour income is due to circumstances and that the extent of inequality of opportunity is higher (0.20) in large urban areas compared to small urban areas and rural areas (0.18 in both cases). This means that in large urban areas 20% of the variation in labour income is due to individual's circumstances rather than effort, against 18% for small urban areas and rural areas. There are considerable differences across NUTS2 regions, with values ranging from 0.14 in Basque Country to 0.36 in Extremadura. The main drivers of such variation, after taking into account differences in life stage, refer to parent's education and main occupation as well as the household financial situation when the individual was aged 14 years old. These results indicate that socio-economic family background acts as a strong condition for individual's income as adults. Similar results were obtained for France, where the level of inequality of opportunity is just over 20%. The degree of inequality of opportunity seems to increase with regional population size, being lower in small urban areas compared to large urban areas, but pronouncedly lower in rural areas. As in Spain, there is large heterogeneity across French regions, with values ranging between 0.10 in Champagne-Ardenne and 0.46 in Haute-Normandie. The analyses carried out do not however explore what may drive such differences between regions. This would be an interesting topic for future research.

Co-funded by the Horizon 2020
Framework Programme of the European Union

4.4. The effect of regional income deprivation on individual income level

In this final part of the empirical analysis using EU-SILC data we implemented longitudinal microeconometric regression models with the aim of measuring the relative importance of NUTS2-level income deprivation on individual labour-income. Of the three EU-SILC countries considered so far it was only possible to measure the effect of NUTS2-level income deprivation on individual income level for Spain and for Finland because there is no regional NUTS2 data for at-risk-of-poverty rates for France.

The main findings from the regression analyses carried out for both countries are that the effect of UTS2-level income deprivation is very small, and sometimes not significant, when considered at such an aggregate level. On the other hand, individual characteristics contribute considerably more to the variation in labour-income observed across individuals. The main finding that at the large NUTS2-level spatial scale the relationship seems to be small, is in line with the national analyses discussed in the following sections.

Co-funded by the Horizon 2020
Framework Programme of the European Union

5. Results from the longitudinal microdata analysis for the United Kingdom

5.1. Introduction

There are not many studies using longitudinal microdata to investigate the relationship between neighbourhood or local area economic disadvantage and individual income outcomes. The most well-known studies for the UK include McCulloch (2001), Buck (2001) and Bolster et al. (2007), who have all used data from the British Household Panel Survey (BHPS) to measure the impact of neighbourhood deprivation on individual outcomes between 1991-1998 (McCulloch, 2001) and 1991-1999 (Buck, 2001, Bolster et al., 2007).[8] Of these studies only Bolster et al. (2007) take advantage of the longitudinal nature of the data by using panel data type regression models. All three studies measure local area disadvantage at the level of wards (for 1991), but Buck (2001) and Bolster et al. (2007) also construct bespoke neighbourhoods by aggregating neighbouring Enumeration Districts according to a series of population thresholds (e.g. 500, 1000, 2000, 5000, 10000). Overall, both McCulloch (2001) and Buck (2001) find evidence of a negative association between neighbourhood disadvantage and individual economic outcomes (e.g. employment, income), while Bolster et al. (2007) conclude there is no evidence supporting a statistically significant relationship. Although Bolster at al. (2007) apply panel data models by differencing out individual effects, there are still considerable limitations with their empirical approach. More specifically, the approach can only control for time-invariant individual unobserved heterogeneity, while at the same time it does not allow estimating the effect of some key individual characteristics on individual outcomes such as educational achievement, ethnicity and family background (all of which are time-invariant and hence drop out of the model). Moreover, to the extent that residential sorting (or any other omitted variables) is associated with time-variant characteristics of individuals, using fixed-effects models or first-difference models will not be sufficient.

[8] There are other relevant studies for the UK but which do not focus specifically on the effect of neighbourhood or area income deprivation on individual economic outcomes, namely, (Clark et al., 2014) and (van Ham and Manley, 2015).

5.2. Discussion of main findings

In this section, we report and discuss the mains results obtained from the regression analyses of local area income deprivation on individuals' labour-income. The reference sample used consists of working-age individuals in full-time employment, the relevant group of individuals for the outcome of interest in the study: work-related labour income. As noted in section 3.2.4, we consider three different types of geographies when measuring low-income deprivation:

1. Fixed boundary and more aggregate administrative geographies based on local authorities or council areas, denoted as LAD;
2. Fixed boundary census-based small area geographies (DZs in the case of Scotland and MSOAs in the case of England) denoted small area; and
3. Bespoke nearest-neighbour geographies, as per Task 5.2, denoted kNN and ranging from 400 to 204800 nearest neighbours.

In addition to considering different types of geographies, we also consider three model specifications. We start with a specification that only includes the measure of local area income deprivation besides having year-specific and country-specific control variables, i.e. model (1). We then add controls for individual demographic and socio-economic characteristics (i.e. gender, age, higher education, industry and occupation), i.e. model (2). The third version of the model specification, that is model (3), includes a control variable for individuals' intention to move and an interaction term between this variable and the measure of local area low-income deprivation[9]. The purpose of model (3) is to investigate if (how) residential satisfaction may influence the results since we expect residential sorting to be one of the channels interfering with the causal effect of residential context on individual outcomes.

[9] Based on the question: "If you could choose, would you stay here in your present home or would you prefer to move somewhere else?".

The results for the models estimated using aggregate administrative geographies for local authorities (i.e. LAD) and the census-based small area geographies (i.e. small area) are reported in Table 1, Table 2, and Table 3 for model (1), model (2), and model (3) respectively. Likewise, the results obtained from the models estimated using bespoke nearest-neighbour geographies (i.e. kNN) are reported in Table 4, Table 5, and Table 6 for model (1), model (2), and model (3) respectively. We report the results for the following estimators: pooled OLS (OLS), random-effects (RE), fixed-effects (FE), and Mundlak's correlated-random effects (CRE). The following paragraphs provide an overview of the main results across model specification and types of geographies.

Looking at the first three tables referring to the aggregate administrative geographies for local authorities (i.e. LAD) and the census-based small area geographies (i.e. small area), we observe that the coefficients are negative in all cases, which indicates that living in areas with higher levels of concentration of low-income households is associated with lower levels of individual labour income. The goodness of fit of the models improves considerably once we take account of individuals' demographic and socio-economic characteristics - i.e. in models (2) and (3) -, with the magnitude of the coefficients of determination increasing to around 40% from less than 10% for model (1).

As expected, the magnitude of the coefficients ranges between the value obtained from the fixed-effects (FE) estimator and that obtained from the pooled OLS. The Hausman test indicates in all cases that the random-effects (RE) estimator is not consistent due to non-zero correlation between the idiosyncratic error term and the covariates. Although the FE estimator produces consistent parameter estimates, the magnitude of the coefficients is likely to be underestimated because it relies only on within-individual variation and makes no use of between-individual variation. Furthermore, the FE estimator is also not very desirable when the variables of interest vary little or slowly over time, which is the case with our measure of local area income deprivation. As an alternative, we therefore also implement the Mundlak correlated random-effects (CRE) estimator, which allows explicitly for correlation between the regressor of interest and the error term by including the mean of the covariate for income deprivation together with the demeaned covariate for income deprivation. We consider the Mundlak CRE to be the preferred estimator from the set of estimators implemented.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Taking the preferred estimator as the reference case, we observe that the association between individual labour income level and local area income deprivation at the more disaggregated level (i.e. small areas) is -0.0173 for model (1) and -0.0108 and -0.0106 for the better-specified model (2) and model (3). This means that an increase of 1 point in the share of low-income households is associated with a reduction of approximately -1.1% in individuals' labour income. On the other hand, if we consider the results obtained for the relationship measured at the level of municipalities (i.e. LAD), the magnitude of the association is -0.019 for model (1) and -0.0124 and -0.0126 for the better-specified model (2) and model (3). Likewise, this means that an increase of 1 point in the share of low-income households is associated with a reduction of approximately -1.2% in individuals' labour income.

One main finding is thus that the size of the negative relationship between residential income deprivation and individual income level is only slightly greater when considered at the LAD level compared to smaller areas approximating neighbourhoods (regardless of the model specification and the estimator used). This indicates that the choice of geographic units and spatial scales matters, even if only mildly. While smaller residential areas may capture better interactions between nearby neighbours living in the same street and the same building, they may not capture well differences in access to key services such as health centres, educational institutions, and other services, which may be located in other parts of the municipality outwith the grounds of the residential neighbourhood.

Moving to the tables focusing on the bespoke nearest-neighbour geographies (i.e. kNN) ranging from 400 to 204800 people, we find overall very similar results. The goodness-of-fit also improves considerable from model (1) to models (2) and (3), and the range of values for the coefficients of local area income deprivation is also bounded by the values produced by the FE and pooled OLS estimators. In consonance to the previous set of tables, we take the Mundlak CRE estimator as the preferred. According to this estimator, the magnitude of the relationship varies between -0.0173 and -0.0178 for kNN=400 and kNN=204800 people and model (1), while for the better-specified models (2) and (3) it ranges from -0.0108 to -0.0130 and from -0.0107 to -0.0127 respectively, for kNN=400 and kNN=204800 people. This shows that the magnitude of the coefficients increases with the number of k-nearest neighbours,

Co-funded by the Horizon 2020
Framework Programme of the European Union

and the values obtained for the different kNN fit very well within the range of values obtained from the small area and LAD regressions.

A final point should be made regarding the specification used in model (3) where we include a control for the intention to move home (i.e. P2M) and an interaction term between this variable and local area income deprivation (i.e. P2M*AROP). The results indicate that the individuals who prefer to move earn on average between -2% to -3% less than those who do not prefer to move. This result may reflect, at least partially, residential sorting on income because higher income individuals are more likely to be able to afford residential locations and houses that meet their preferences. However, and perhaps not intuitively, the interaction between preference to move and local area income deprivation is positive, suggesting that the negative association between area deprivation and individual income is slightly smaller for those who prefer to move; we also observe that the statistical significance of the interaction term tends to reduce with the size/spatial scale of the geographical units.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 1: Regressions of individual labour-income on LAD and small area low-income deprivation – Model (1)

Variables	Model (1) - Pooled OLS		Model (1) - RE		Model (1) - FE		Model (1) - CRE	
	Small area[1]	LAD[2]	Small area[1]	LAD[2]	Small area[1]	LAD[2]	Small area[1]	LAD[2]
AROP	-0.0206***	-0.0240***	-0.0129***	-0.0180***	-0.0025***	-0.0048***	-0.0027***	-0.0050***
mean AROP	-	-	-	-	-	-	-0.0173***	-0.0191***
Prefer to move (P2M)	-	-	-	-	-	-	-	-
AROP*Prefer to move (P2M)	-	-	-	-	-	-	-	-
Hausman test (FE vs RE)	-	-	1062.38***	684.8***	-	-	-	-
Controls for year	YES	YES	YES	YES	YES	YES	YES	YES
Controls for country	YES	YES	YES	YES	YES	YES	YES	YES
Controls for individual attributes	NO	NO	NO	NO	NO	NO	NO	NO
Preference to move (Yes/No)	NO	NO	NO	NO	NO	NO	NO	NO
Observations	80,179	80,179	80,179	80,179	80,179	80,179	80,179	80,179
Adjusted R^2	0.07	0.05	-	-	-	-	-	-
R^2 - overall	-	-	0.06	0.05	0.03	0.03	0.07	0.05
R^2 -within	-	-	0.17	0.17	0.18	0.18	0.18	0.18
R^2 -between	-	-	0.03	0.02	0.01	0.00	0.04	0.02

*, **, *** denote statistical significance p<0.10, p<0.05, p<0.01 respectively.

[1] small areas refer to data zones/LSOAs in Scotland and MSOAs in England

[2] LAD correspond to Local Authority Districts.

The sample consists of full-timers in paid employment.

Table 2: Regressions of individual labour-income on LAD and small area low-income deprivation – Model (2)

Variables	**Model (2) - Pooled OLS**		**Model (2) - RE**		**Model (2) - FE**		**Model (2) - CRE**	
	Small area[1]	**LAD[2]**	**Small area[1]**	**LAD[2]**	**Small area[1]**	**LAD[2]**	**Small area[1]**	**LAD[2]**
AROP	-0.0111***	-0.0144***	-0.0095***	-0.0137***	-0.0021***	-0.0042***	-0.0020***	-0.0038***
mean AROP	-	-	-	-	-	-	-0.0108***	-0.0126***
Prefer to move (P2M)	-	-	-	-	-	-	-	-
AROP*Prefer to move (P2M)	-	-	-	-	-	-	-	-
Hausman test (FE vs RE)	-	-	4430.09***	4604.22***	-	-	-	-
Controls for year	YES	YES	YES	YES	YES	YES	YES	YES
Controls for country	YES	YES	YES	YES	YES	YES	YES	YES
Controls for individual attributes	YES	YES	YES	YES	YES	YES	YES	YES
Preference to move (Yes/No)	NO	NO	NO	NO	NO	NO	NO	NO
Observations	78,612	78,612	78,612	78,612	78,612	78,612	78,612	78,612
Adjusted R^2	0.42	0.42	-	-	-	-	-	-
R^2 - overall	-	-	0.41	0.40	0.09	0.09	0.41	0.40
R^2 -within	-	-	0.16	0.16	0.19	0.19	0.16	0.16
R^2 -between	-	-	0.42	0.42	0.07	0.07	0.42	0.41

*, **, *** denote statistical significance p<0.10, p<0.05, p<0.01 respectively.

[1] small areas refer to data zones/LSOAs in Scotland and MSOAs in England

[2] LAD correspond to Local Authority Districts.

The sample consists of full-timers in paid employment.

Table 3: *Regressions of individual labour-income on LAD and small area low-income deprivation – Model (3)*

Variables	Model (3) - Pooled OLS		Model (3) - RE		Model (3) - FE		Model (3) - CRE	
	Small area[1]	LAD[2]	Small area[1]	LAD[2]	Small area[1]	LAD[2]	Small area[1]	LAD[2]
AROP	-0.0123***	-0.0147***	-0.0102***	-0.0142***	-0.0024***	-0.0046***	-0.0026***	-0.0044***
mean AROP	-	-	-	-	-	-	-0.0106***	-0.0124***
Prefer to move (P2M)	-0.0605***	-0.0344*	-0.0288***	-0.0251**	-0.0091	-0.0116	-0.0249***	-0.0225**
AROP*Prefer to move (P2M)	0.0027***	0.0008	0.0015***	0.0012*	0.0005	0.0007	0.0012***	0.0010
Hausman test (FE vs RE)	-	-	4466.69***	4531.68***	-	-	-	-
Controls for year	YES	YES	YES	YES	YES	YES	YES	YES
Controls for country	YES	YES	YES	YES	YES	YES	YES	YES
Controls for individual attributes	YES	YES	YES	YES	YES	YES	YES	YES
Preference to move (Yes/No)	YES	YES	YES	YES	YES	YES	YES	YES
Observations	78,191	78,191	78,191	78,191	78,191	78,191	78,191	78,191
Adjusted R^2	0.42	0.42	-	-	-	-	-	-
R^2 - overall	-	-	0.41	0.40	0.09	0.09	0.41	0.40
R^2 -within	-	-	0.16	0.16	0.19	0.19	0.16	0.16
R^2 -between	-	-	0.42	0.42	0.07	0.07	0.42	0.42

*, **, *** denote statistical significance p<0.10, p<0.05, p<0.01 respectively.
The sample consists of full-timers in paid employment.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 4: Regressions of individual labour-income on kNN bespoke neighbourhood low-income deprivation – model (1)

Bespoke geographies	Pooled OLS	RE	FE	CRE	
	AROP	AROP	AROP	AROP	mean AROP
kNN = 400	-0.0205***	-0.0129***	-0.0025***	-0.0026***	-0.0173***
kNN = 800	-0.0205***	-0.0129***	-0.0025***	-0.0027***	-0.0173***
kNN = 1600	-0.0205***	-0.0129***	-0.0025***	-0.0027***	-0.0172***
kNN = 3200	-0.0210***	-0.0136***	-0.0030***	-0.0032***	-0.0173***
kNN = 6400	-0.0214***	-0.0142***	-0.0031***	-0.0033***	-0.0177***
kNN = 12800	-0.0230***	-0.0157***	-0.0037***	-0.0040***	-0.0184***
kNN = 25600	-0.0239***	-0.0171***	-0.0042***	-0.0045***	-0.0188***
kNN = 51200	-0.0242***	-0.0182***	-0.0052***	-0.0055***	-0.0182***
kNN = 102400	-0.0245***	-0.0192***	-0.0057***	-0.0060***	-0.0180***
kNN = 204800	-0.0256***	-0.0212***	-0.0069***	-0.0075***	-0.0178***
Hausman test (FE vs RE)	-	rejects H_0 for all kNN		-	
Controls for year	YES	YES	YES	YES	
Controls for country	YES	YES	YES	YES	
Controls for individual attributes	NO	NO	NO	NO	
Preference to move (Yes/No)	NO	NO	NO	NO	
Observations	80,179	80,179	80,179	80,179	
Adjusted R^2 (range)	0.05-0.07	-	-	-	
R^2 - overall (range)	-	0.04-0.06	0.03	0.05-0.07	
R^2 -within (range)	-	0.17	0.18	0.18	
R^2 -between (range)	-	0.02-0.03	0.01	0.02-0.04	

*, **, *** denote statistical significance p<0.10, p<0.05, p<0.01 respectively.

The sample consists of full-timers in paid employment.

Table 5: Regressions of individual labour-income on kNN bespoke neighbourhood low-income deprivation – model (2)

Bespoke geographies	Pooled OLS	RE	FE	CRE	
	AROP	AROP	AROP	AROP	mean AROP
kNN = 400	-0.0111***	-0.0095***	-0.0020***	-0.0018***	-0.0108***
kNN = 800	-0.0111***	-0.0095***	-0.0020***	-0.0019***	-0.0108***
kNN = 1600	-0.0111***	-0.0095***	-0.0020***	-0.0018***	-0.0109***
kNN = 3200	-0.0117***	-0.0102***	-0.0026***	-0.0024***	-0.0109***
kNN = 6400	-0.0123***	-0.0108***	-0.0026***	-0.0024***	-0.0116***
kNN = 12800	-0.0134***	-0.0119***	-0.0031***	-0.0030***	-0.0119***
kNN = 25600	-0.0143***	-0.0131***	-0.0037***	-0.0036***	-0.0123***
kNN = 51200	-0.0147***	-0.0139***	-0.0045***	-0.0042***	-0.0122***
kNN = 102400	-0.0154***	-0.0146***	-0.0046***	-0.0040***	-0.0128***
kNN = 204800	-0.0163***	-0.0158***	-0.0054**	-0.0046***	-0.0130***
Hausman test (FE vs RE)	-	rejects H_0 for all kNN		-	
Controls for year	YES	YES	YES	YES	
Controls for country	YES	YES	YES	YES	
Controls for individual attributes	YES	YES	YES	YES	
Preference to move (Yes/No)	NO	NO	NO	NO	
Observations	78,612	78,612	78,612	78,612	
Adjusted R^2 (range)	0.42	-	-	-	
R^2 - overall (range)	-	0.40-0.41	0.09	0.40-0.41	
R^2 -within (range)	-	0.16	0.19	0.16	
R^2 -between (range)	-	0.42	0.07	0.41-0.42	

*, **, *** denote statistical significance p<0.10, p<0.05, p<0.01 respectively.
The sample consists of full-timers in paid employment.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 6: Regressions of individual labour-income on kNN bespoke neighbourhood low-income deprivation – model (3)

Bespoke geographies	**Pooled OLS**			**RE**			**FE**			**CRE**			
	AROP	**P2M**	**AROP*P2M**	**AROP**	**P2M**	**AROP*P2M**	**AROP**	**P2M**	**AROP*P2M**	**AROP**	**mean AROP**	**P2M**	**AROP*P2M**
kNN = 400	-0.0122***	-0.0597***	0.0026***	-0.0102***	-0.0287***	0.0015***	-0.0023***	-0.0089	0.0005	-0.0025***	-0.0107***	-0.0247***	0.0012***
kNN = 800	-0.0123***	-0.0599***	0.0027***	-0.0102***	-0.0281***	0.0015***	-0.0023***	-0.0083	0.0005	-0.0025***	-0.0107***	-0.0243***	0.0012***
kNN = 1600	-0.0123***	-0.0605***	0.0027***	-0.0102***	-0.0281***	0.0014***	-0.0023***	-0.0081	0.0005	-0.0025***	-0.0107***	-0.0241***	0.0012***
kNN = 3200	-0.0128***	-0.0610***	0.0027***	-0.0107***	-0.0251***	0.0012***	-0.0027***	-0.0033	0.0002	-0.0029***	-0.0108***	-0.0212***	0.0010**
kNN = 6400	-0.0134***	-0.0603***	0.0026***	-0.0114***	-0.0280***	0.0014***	-0.0028***	-0.0071	0.0004	-0.0030***	-0.0114***	-0.0235***	0.0011**
kNN = 12800	-0.0144***	-0.0613***	0.0026***	-0.0126***	-0.0316***	0.0016***	-0.0035***	-0.0100	0.0006	-0.0037***	-0.0118***	-0.0277***	0.0013***
kNN = 25600	-0.0152***	-0.0564***	0.0022***	-0.0138***	-0.0332***	0.0017***	-0.0042***	-0.0128	0.0007	-0.0043***	-0.0121***	-0.0289***	0.0014***
kNN = 51200	-0.0156***	-0.0581***	0.0023**	-0.0146***	-0.0351***	0.0018***	-0.0050***	-0.0150	0.0009	-0.0050***	-0.0120***	-0.0314***	0.0015***
kNN = 102400	-0.0162***	-0.0539***	0.0020*	-0.0153***	-0.0339***	0.0017***	-0.0052***	-0.0165	0.0010	-0.0049***	-0.0125***	-0.0307***	0.0015**
kNN = 204800	-0.0169***	-0.0482**	0.0016	-0.0164***	-0.0311**	0.0015**	-0.0060**	-0.0156	0.0009	-0.0054***	-0.0127***	-0.0283**	0.0013*
Hausman test (FE vs RE)	-			rejects H0 for all kNN			-			-			
Controls for year	YES			YES			YES			YES			
Controls for country	YES			YES			YES			YES			
Controls for individual X	YES			YES			YES			YES			
Preference to move	YES			YES			YES			YES			
Observations	78,191			78,191			78,191			78,191			
Adjusted R^2 (range)	0.42			-			-			-			
R^2 - overall (range)	-			0.10-0.41			0.09			0.40-0.41			
R^2 -within (range)	-			0.16			0.19			0.16			
R^2 -between (range)	-			0.42			0.07			0.42			

*, **, *** denote statistical significance p<0.10, p<0.05, p<0.01 respectively. P2M= prefers to move home.
The sample consists of full-timers in paid employment.

5.3. Conclusion

Overall, the results indicate that local area income deprivation affects individual labour-income level, that is, higher concentration of poor households is associated with lower individual income. However, the relative importance of this area effect is much smaller than the contribution of individual's characteristics such as educational attainment and occupation.

The magnitude of the effect varies depending of the spatial scale and type of geography, but only marginally. For bespoke neighbourhoods ranging between 400 people and 204,800 people, the impact of increasing the at-risk-of-poverty rate by 1 percentage point on income level ranges between -1.1%an d-1.3% respectively. For administrative or statistical geographies based on LSOAs/MSOAs and LADs, we obtain marginal effects ranging between -1.1% and -1.2% respectively. Consequently, the results do not indicate strong differences across spatial scales.

It is worth noting that the UK-based empirical analyses only measure the contemporary effect of residential area income deprivation on individual's income, and thus are likely to suffer from simultaneity bias and self-selection on time variant unobservables. Consequently, the findings do not establish causal relations, but rather conditional associations.

6. Results from the longitudinal microdata analysis for Sweden

6.1. Introduction

The availability and access to register-based geocoded microdata in Sweden has allowed researchers to implement more sophisticated econometric analyses of area or neighbourhood effects on individual income, which can better address issues relating to residential self-selection and the choice of the spatial units of relevance. Recent examples of such studies focusing on individual income outcomes include Brännström (2005), Andersson et al. (2007), Galster et al. (2008), Hedman and Galster (2013), Hedman et al. (2015), Andersson and Malmberg (2016), Mellander et al. (2017), and Wimark et al. (2019).

Brännström (2005) uses longitudinal data from the Stockholm Birth Cohort Study and a multilevel modelling approach to estimate the importance of neighbourhood factors experienced in childhood, adolescence and young adulthood on income level of individuals in later life as adults for a cohort born in 1953. The size of the neighbourhood effect is given by the magnitude of variance partition coefficient (VPC) associated with two spatial scales: census tracts and parishes. The degree of variation in individual income accounted for by the two spatial units (census and parish) measures the importance of the contextual effects experienced by individuals when they were growing up. The results from the longitudinal multilevel models indicate that prior residential location accounts only for a very little proportion of the variation in individuals' income later in life, suggesting a very modest, if any, role for neighbourhood effects.

Andersson et al. (2007) used Swedish longitudinal microdata to investigate the effect of multiple neighbourhood factors, notably the concentration of income poverty measured by the share of adult males with earnings in the lowest 30th percentile, on individual earnings living in Sweden between 1996 and 1999 (no exposure time). Residential location is operationalised using three geographies: small area market statistics (SAMS) with an average population of 1000 people, the municipality of residence and associated local labour market. Concentration of low income is measured at the SAMS level. In spite of using longitudinal microdata, the authors do not implement panel data type estimators and limit their analysis to the simpler OLS estimator. The authors find effects from living in deprived areas and therefore suggest a continuation of area-based programmes. The above cited Brännström

(2005) on the other hand do not recommend such programmes, instead he endorses welfare policies on a general level.

Galster et al. (2008) use register-based Swedish longitudinal microdata for working-age adults living in metropolitan Sweden for the period between 1991 and 1999 to estimate the effect of neighbourhood income mix on individuals’ income from work. Similar to other Swedish work, they define neighbourhoods based on the small area market statistics (SAMS) defined by Statistics Sweden. SAMS are designed based on information about housing type, tenure and construction period in order to identify which are relatively homogenous. The indicator of income mix, at SAMS level, consists of the share of working age males in the lowest 30% of the income distribution, the highest 30% of the income distribution, and the middle 40%. To control for potential unobserved time-invariant individual heterogeneity, they estimate first-difference models of changes in average income between 1991-1995 and 1996-1999. In addition, to control for potential unobserved time-varying characteristics they re-estimate the models using a sample of non-movers. Compared to simple OLS, these two methods produce lower but still statistically significant neighbourhood effects for the three income indicators.

Hedman and Galster (2013) use Swedish longitudinal microdata for males residing in Stockholm metropolitan area from 1994 to 2006 to measure the effect of neighbourhood income mix on individual earnings. They develop panel data regression models which combine individual fixed-effects (FE) with instrumental variables (FE-IV) to control for neighbourhood selection and correct for simultaneity bias between neighbourhood selection and individual income outcome. The instruments used for individual income, which affect individual income but not neighbourhood income mix, binary variables that measure whether an individual was: on sick leave during a given year, on parental leave during a given year, and/or receiving pre-retirement benefits. The instruments for neighbourhood income, which affect neighbourhood income but not individual income, include: individual–partner ethnic combination, individual–partner ethnic combination interacted by number of children, individual–partner ethnic combination interacted with partner income, and the share of males in each of three age groups of children in the household. They find that the magnitude of neighbourhood income mix increases once neighbourhood selection and endogeneity are accounted for using the IV-FE estimator. Neighbourhoods are defined as in the previous study by Galster et al. (2008) using SAMS and the indicator of income mix consists of the share of working age males in the lowest 30% of the income distribution, the highest 30% of the income distribution, and the middle 40%.

The aim of the analysis by Andersson and Malmberg (2016) was to examine how poverty risks and early income career at adult age are influenced by different neighbourhood contexts in early youth. They again use Swedish longitudinal register data and follow individuals, in this study a cohort born in 1980 and follow them until year 2012. Their residential context is defined in 1995 when the cohort is 15 and measured by expanding a buffer around the residential locations of each individual and, by computing statistical aggregates of different sociodemographic variables for that population using Equipop (Östh et al., 2014). Their results show that poverty risks increase for individuals growing up in areas characterised by high numbers of social assistance recipients living nearby, whereas 'elite' geographical context is advantageous for both women's and men's future income. The study represents a national comprehensive study unlike the other studies with Stockholm or the three metropolitan areas.

Mellander et al. (2017) investigate the relative importance of residential vs workplace neighbourhood effects on working age adults' income based on Swedish longitudinal microdata for the period 2002-2011. They measure neighbourhood effects as the concentration of individuals in low- and high-skill occupations at different scales, namely: 250 x 250 m and 1000 x 1000m blocks for urban and non-urban locations respectively, SAMS, municipalities, and local labour markets. The magnitude of the residence-based neighbourhood effects is larger for blocks and SAMS but considerably lower for municipalities and local labour markets. On the other hand, the relative magnitude of workplace-based neighbourhood effects tend to be larger for SAMS and municipalities, followed by local labour markets. In spite of using longitudinal data the analyses are based on simple OLS estimators, which cannot correct for the main estimation issues discussed earlier in previous sections of the report.

A recent study by Wimark et al. (2019) investigate the income as an effect of initial settlement for migrant's income. They use five variables, employment, education, migrant, income and, social welfare on three different number of neighbours k=200, 2000 and 20000 in 1996, 2002 and 2008. All the indicators are used in a factor analysis to determine affluent and deprived areas including scale. Migrants arriving the year before estimation and was found to be negatively affected from living in a deprived neighbourhood concerning their income and employment. Stronger positive effects were however estimated for residing in an affluent area when arriving to Sweden.

Of the studies mentioned above, Andersson and Malmberg (2016), Wimark et al. (2019) as well as Hedman et al. (2015) constructed ego-centric bespoke neighbourhoods using a k-nearest neighbour and the software Equipop. The analyses we implemented in Task 5.2 of Work Package 5 followed the same approach for a series of k-nearest neighbours (e.g. 200, 400, 800, 1600, …, 204800). Furthermore, and unlike the majority of the studies above, the Hedman et al. (2015) implemented panel data type estimators which allow correcting, even if only partially, for some of the estimation issues mentioned earlier, notably individual unobserved heterogeneity correlated with residential self-selection and omitted variable bias.

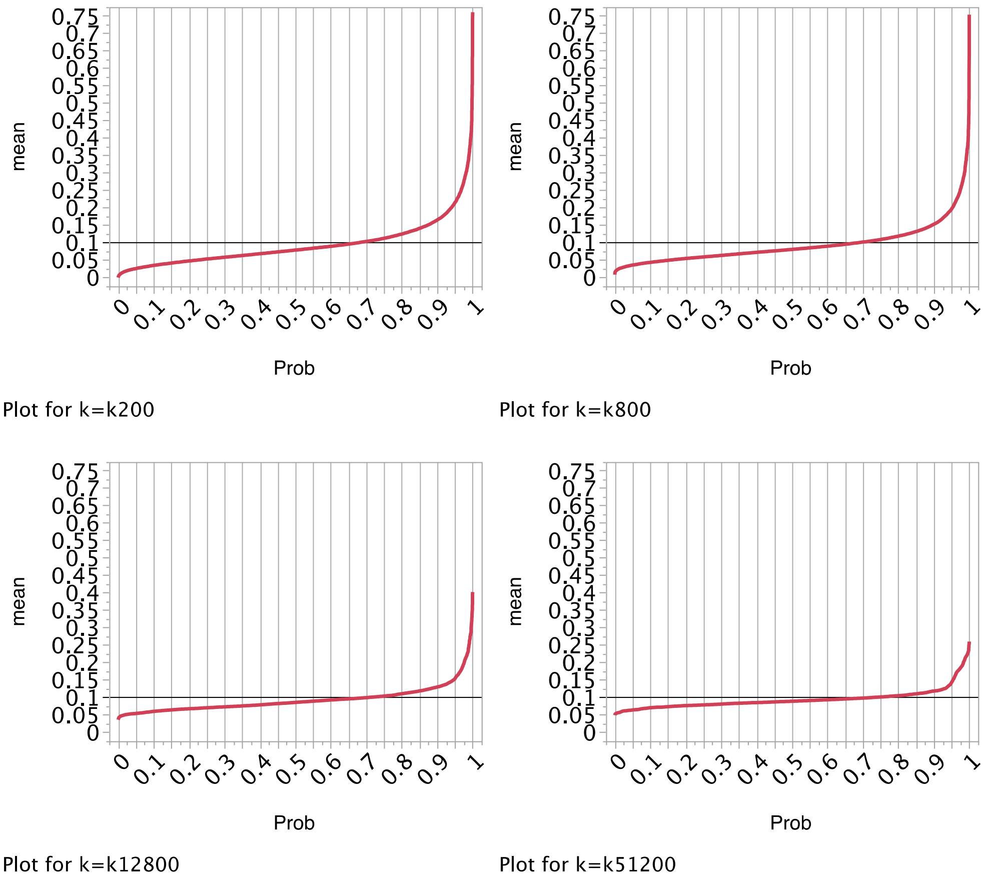
6.2. Discussion of main findings

In order to analyse local area income deprivation on individuals' labour-income later in life we have used regression analyses in the following section on Swedish data. A cohort born in 1986 is used to test if earlier life surroundings influence later life labour-income. The particular interest in this study is if income deprivation, or poverty, in the local or larger area during upbringing influence individuals' futures in the form of income in late 20s. The table below provides descriptive statistics for the individual characteristics.

Table 7: Descriptives

Variable	Obs.	Mean	Std. Dev.	Min	Max
Income age 25-30 (max percentile)	103,945	67,128	25,329	4,664	100
Male	103,945	0,515	0,500	0	1
Non-European background	103,945	0,083	0,276	0	1
Single mother 2001	103,945	0,217	0,412	0	1
Social assistance 2001	103,945	0,073	0,261	0	1
Tertiary education 2001	103,945	0,555	0,738	0	2
Disposable income per parent 2001 (€ 1000s)	103,945	10,3	14,4	0	2809,0
Employment 2001	103,945	1,489	0,657	0	2,000
IntervalRatio_200	103,945	0,090	0,074	0,000	0,752
IntervalRatio_1600	103,945	0,088	0,059	0,010	0,726
IntervalRatio_12800	103,945	0,087	0,039	0,035	0,398
IntervalRatio_51200	103,945	0,087	0,028	0,046	0,246
IntervalRatio_204800	103,945	0,088	0,015	0,060	0,136
Neighborhood_percentile_k_200	103,945	41,298	29,910	0,048	10,000
Neighborhood_percentile_k_1600	103,945	41,820	30,536	0,000	10,000
Neighborhood_percentile_k_12800	103,945	46,020	30,646	0,000	10,000
Neighborhood_percentile_k_51200	103,945	49,656	30,082	0,003	10,000
Neighborhood_percentile_k_204800	103,945	50,309	29,287	0,001	9,999
NUTS2 mean poor	103,945	0,140	0,012	0,118	0,162
NUTS3 mean poor (County)	103,945	0,140	0,014	0,109	0,165

Figure 1 shows percentile plots for the proportion of individuals, aged 25 or older, that are at risk of poverty (disposable income less than 60% of the median) for individualized neighbourhoods defined using different k-levels (200, 800, 12800 and 51200). These plots demonstrate that the at-risk-of-poverty population is segregated. Most neighbourhoods have relative low proportions of at-risk-of-poverty individuals (y-axis) but in a small share of the neighbourhoods the concentration of at-risk-of-poverty individuals is very high (right-hand side of graphs). For example, the graph of neighbourhoods of 12800 individuals show rather modest proportions of poor (varying under 10%) up to the 75th percentile, that is ¾ of neighbourhoods. The last ¼ of neighbourhoods have larger proportions of poor, ranging from 10 to 40 percent poor. This pattern of variation suggests that models using percentiles to measure neighbourhood context will not capture well the effect of varying poverty rates.



Plot for k=k200

Plot for k=k800

Plot for k=k12800

Plot for k=k51200

Figure 1. Percentile plots for the proportion of individuals at risk of poverty

Notes: AROP refers to individuals, aged 25 or older, that are at risk of poverty (disposable income less than 60% of the median) for individualized neighbourhoods defined using different k-levels.

Table 8 shows how different subgroups are sorted into low poverty and high poverty neighbourhoods. The table demonstrate that there is sorting into poor and non-poor neighbourhoods also with respect to other characteristics of the individuals in our sample. Thus, individuals with non-European parents, with parents without employment and from families with social assistance are strongly overrepresented in poor neighbourhoods. Whereas individuals whose parents have a tertiary education or are Swedish born are overrepresented in non-poor neighbourhoods. The same is true for individuals who come from households that do not have social assistance or non-single mother families. These patterns are most evident for small scale neighbourhoods, but also for the largest contexts (k=204,800) there is a strong overrepresentation of individuals with non-European parents in the poorest neighbourhoods.

Table 8: Distribution of subgroups across low and high poverty neighbourhoods. Neighbourhoods ranging from first to 10th percentile (0-9).

k-value 200

	Non-European parents		Single mother household		Family has social allowance		Employed parents			Parents with teriary education		
Poverty	No	Yes	No	Yes	No	Yes	None	1	2	None	1	2
0	18,3%	6,4%	18,9%	11,8%	18,3%	4,1%	5,2%	13,2%	21,5%	13,8%	20,1%	26,2%
1	14,9%	6,4%	14,7%	12,5%	14,8%	6,2%	7,3%	13,3%	15,8%	12,4%	16,0%	18,4%
2	12,8%	7,1%	12,3%	12,4%	12,7%	7,2%	8,3%	12,3%	13,0%	11,5%	13,4%	13,9%
3	11,1%	6,9%	10,5%	11,6%	11,0%	7,5%	9,0%	11,4%	10,7%	10,6%	11,1%	10,8%
4	9,9%	6,8%	9,2%	11,0%	9,7%	8,3%	9,1%	10,6%	9,2%	9,8%	9,6%	8,9%
5	8,0%	6,1%	7,6%	8,7%	7,9%	7,9%	7,8%	8,7%	7,4%	8,4%	7,6%	6,1%
6	6,8%	6,3%	6,6%	7,4%	6,7%	7,8%	8,2%	7,2%	6,3%	7,6%	6,0%	4,7%
7	6,1%	6,3%	5,9%	6,7%	5,9%	8,2%	7,9%	6,5%	5,6%	7,1%	5,1%	3,9%
8	5,3%	8,4%	5,5%	5,9%	5,3%	9,0%	8,3%	6,2%	4,8%	6,7%	4,7%	2,9%
9	6,8%	39,3%	8,8%	12,0%	7,6%	33,8%	28,7%	10,8%	5,7%	12,1%	6,4%	4,3%

k-value 1600

	Non-European parents		Single mother household		Family has social allowance		Employed parents			Parents with teriary education		
Poverty	No	Yes	No	Yes	No	Yes	None	1	2	None	1	2
0	18,8%	7,4%	18,9%	14,4%	18,8%	5,7%	7,6%	15,3%	20,9%	14,0%	21,2%	27,5%
1	14,9%	7,5%	14,5%	13,7%	14,8%	8,2%	9,0%	13,9%	15,4%	13,1%	15,7%	16,8%
2	12,2%	7,2%	11,7%	12,1%	12,1%	8,2%	9,2%	12,0%	12,1%	11,0%	12,8%	13,2%
3	10,5%	6,9%	10,1%	10,7%	10,3%	9,3%	9,1%	10,4%	10,3%	10,3%	10,2%	10,1%
4	9,2%	5,9%	8,8%	9,4%	9,0%	8,3%	8,6%	9,3%	8,8%	9,3%	8,9%	7,7%
5	8,0%	5,9%	7,7%	8,0%	7,9%	6,7%	7,4%	8,1%	7,7%	8,3%	7,5%	6,4%
6	7,0%	5,9%	6,8%	7,1%	6,8%	7,5%	7,8%	7,0%	6,7%	7,6%	6,3%	4,9%
7	6,1%	5,2%	6,1%	5,8%	6,0%	6,4%	6,3%	6,1%	5,9%	6,7%	5,4%	4,3%
8	5,6%	6,5%	5,7%	5,7%	5,5%	7,7%	7,3%	5,9%	5,3%	6,7%	4,5%	3,7%
9	7,6%	41,5%	9,7%	13,1%	8,7%	31,8%	27,7%	11,9%	6,9%	13,0%	7,6%	5,3%

k-value 12800

	Non-European parents		Single mother household		Family has social allowance		Employed parents			Parents with teriary education		
Poverty	No	Yes	No	Yes	No	Yes	None	1	2	None	1	2
0	14,8%	8,3%	14,4%	13,6%	14,7%	8,6%	9,7%	13,6%	15,3%	11,5%	16,8%	20,8%
1	13,4%	8,1%	12,9%	13,1%	13,2%	9,8%	10,1%	13,0%	13,3%	12,1%	13,7%	14,9%
2	11,5%	7,0%	11,1%	11,2%	11,3%	9,0%	8,9%	11,3%	11,4%	10,8%	11,6%	11,8%
3	9,6%	5,7%	9,5%	8,5%	9,4%	6,8%	7,2%	8,8%	9,8%	9,0%	9,6%	9,4%
4	9,7%	6,0%	9,5%	9,1%	9,5%	7,6%	7,8%	9,2%	9,7%	9,4%	9,7%	8,6%
5	8,3%	7,0%	8,2%	8,3%	8,2%	7,7%	7,6%	8,3%	8,2%	8,4%	8,1%	7,4%
6	7,9%	6,6%	7,9%	7,5%	7,8%	7,6%	7,2%	7,8%	7,9%	8,5%	7,1%	6,3%
7	8,0%	9,6%	8,2%	7,9%	8,1%	8,7%	9,2%	8,2%	8,0%	8,8%	7,4%	7,0%
8	7,5%	6,9%	7,5%	7,3%	7,4%	8,1%	8,4%	7,5%	7,3%	8,1%	6,6%	6,1%
9	9,3%	34,8%	10,9%	13,5%	10,3%	26,1%	24,0%	12,4%	9,0%	13,4%	9,3%	7,6%

k-value 51200

	Non-European parents		Single mother household		Family has social allowance		Employed parents			Parents with teriary education		
Poverty	No	Yes	No	Yes	No	Yes	None	1	2	None	1	2
0	11,8%	7,3%	11,5%	10,8%	11,6%	8,2%	8,5%	11,0%	12,1%	10,0%	12,7%	14,6%
1	10,4%	7,1%	10,2%	9,7%	10,2%	8,8%	8,3%	9,8%	10,5%	9,2%	10,8%	12,6%
2	10,1%	7,2%	9,9%	9,6%	9,9%	8,1%	8,2%	9,7%	10,2%	9,8%	10,0%	9,6%
3	10,6%	8,9%	10,3%	10,7%	10,5%	10,1%	9,3%	10,5%	10,5%	10,6%	10,3%	10,1%
4	10,2%	8,6%	10,0%	10,2%	10,2%	8,5%	9,7%	10,4%	9,9%	9,8%	10,4%	10,5%
5	9,8%	6,8%	9,6%	9,7%	9,7%	8,0%	8,5%	9,8%	9,7%	9,6%	9,8%	9,4%
6	9,5%	7,9%	9,4%	9,2%	9,3%	9,9%	9,7%	9,1%	9,4%	9,6%	9,1%	8,9%
7	8,0%	7,1%	8,1%	7,4%	8,0%	7,8%	8,1%	7,6%	8,1%	8,6%	7,4%	6,3%
8	8,8%	6,9%	8,7%	8,2%	8,6%	8,4%	8,3%	8,5%	8,7%	9,1%	8,1%	7,4%
9	10,9%	32,1%	12,2%	14,4%	11,9%	22,2%	21,5%	13,6%	10,8%	13,7%	11,5%	10,7%

k-value 204800

	Non-European parents		Single mother household		Family has social allowance		Employed parents			Parents with teriary education		
Poverty	No	Yes	No	Yes	No	Yes	None	1	2	None	1	2
0	11,1%	4,4%	10,7%	9,7%	10,7%	8,1%	8,5%	10,3%	10,9%	10,1%	11,2%	11,0%
1	10,8%	6,2%	10,5%	10,0%	10,5%	9,3%	8,6%	10,4%	10,6%	10,5%	10,2%	10,0%
2	9,8%	9,2%	9,7%	10,1%	9,7%	11,0%	11,0%	10,0%	9,5%	9,8%	9,6%	10,0%
3	8,8%	7,7%	8,5%	9,2%	8,7%	8,0%	7,7%	8,9%	8,7%	7,8%	9,4%	10,8%
4	9,7%	7,3%	9,7%	8,9%	9,6%	8,4%	7,8%	9,2%	10,0%	9,8%	9,4%	8,7%
5	10,0%	10,5%	10,1%	9,8%	10,1%	9,6%	9,3%	9,8%	10,4%	10,0%	10,3%	10,0%
6	10,1%	8,1%	10,0%	9,7%	10,1%	8,1%	8,9%	9,7%	10,2%	10,3%	9,7%	8,9%
7	10,4%	6,4%	10,1%	9,8%	10,1%	9,2%	10,2%	10,2%	9,9%	10,8%	9,4%	8,0%
8	10,4%	14,5%	10,8%	10,7%	10,7%	11,2%	10,9%	10,5%	10,9%	10,5%	10,9%	11,6%
9	8,9%	25,9%	9,8%	12,0%	9,8%	17,0%	17,1%	11,0%	8,9%	10,3%	9,9%	11,0%

As was seen from the description above the proportion of poor varies between neighbourhoods and geographical scales. The next step in our analyses was to find out if individuals were differently affected while inhabiting areas with a certain proportion of poor during adolescence.

Our first model (model 1) was run with controls only, individual and family characteristics, see Table 9. As can be seen being male is associated with a positive estimate on a high labour income later in life equivalent to more than nine percentiles. Thus being a man is the most important characteristic for having a high income in our population. A positive effect on income is also true for individuals having one or two parents with tertiary education and with parents with higher incomes. Which is in line with earlier results. Positive estimates for later incomes were found for individuals with one or two employed parents too. A negative association with later life income was found for individuals with one or two parents born outside Europe, individuals living with a single mother (not large estimate) and lastly but most importantly a negative effect was found for individuals in households with social assistance.

Table 9: Model with individual and family characteristics explaining income in young adulthood.

	(model 1) Income	Std Error
Male	9.444***	.152
Non-European background	-1.086***	.291
Single mother 2001	-.945***	.233
Social assistance 2001	-6.015***	.333
Tertiary education 2001 none		.
Tertiary education 2001 one	1.267***	.181
Tertiary education 2001 two	.912***	.231
Disposable income 2001, Euro 1000s	0.026***	.005
Employment 2001 none		.
Employment 2001 one	4.458***	.310
Employment 2001 two	8.628***	.337
Cons	55.804***	.339
N	103945	
adj. R^2	.062	

Standard errors in second column. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In models 2 to model 6 in Table 10, the ratio of income poor in the area is added to the test for the outcome labour-income in later life. The ratio of income poor is tested for differently sized surrounding populations starting with the proportion of poor among the 200 closest neighbours and ranging to the model no. 6 with the ratio of poor among the 204,800 closest inhabitants. The individual and family controls are largely unaffected by the introduction of the neighbourhood levels (same order of magnitude and signs and not included). In sum the effect on this cohorts' future income is negative the larger the share of poor residents in the neighbourhood and larger surrounding geographical area. As the ratio has different standard deviations across neighbourhood sizes (the larger the population the larger the Std.) we have also made an analysis where neighbourhoods are ranked from richest to poorest in percentiles, see Table 11.

Table 10: Models with differently sized contexts, poverty rates. Controls as in Table 1.

	(2) Income	Std Error	(3) Income	Std Error	(4) Income	Std Error	(5) Income	Std Error	(6) Income	Std Error
Poverty rate k=200	-11.697***	1.147								
Poverty rate k=1600			-14.270***	1.407						
Poverty rate k=12800					-12.784***	2.032				
Poverty rate k=51200							-14.231***	2.812		
Poverty rate k=204800									-17.265***	5.065
cons	57.224***	.367	57.343***	.372	57.027***	.391	57.106***	.426	57.328***	.561
N	103945		103945		103945		103945		103945	
adj. R^2	.063		.063		.063		.063		.062	

Standard errors in second column. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 11 shows effect sizes for different neighbourhood percentiles. They suggest that we find the strongest effects at the k=1600 level from the ratio of poor in the neighbourhood. That is, living in the five percent poorest areas the negative effect on future income ranges from -2.59 to -5.26 percentiles in income. A neighbourhood of 1600 inhabitants in the Swedish context can be considered fairly common place and can be found in both smaller cities and larger metropolitan areas.

Table 11: Effect sizes on young adult income percentile from neighbourhood context.

Neighbourhood percentile	**k200**	**k1600**	**k12800**	**k51200**
0.1	-0.40	-0.60	-0.77	-1.00
0.25	-0.62	-0.77	-0.90	-1.12
0.5	-0.92	-1.04	-1.10	-1.27
0.75	-1.31	-1.42	-1.33	-1.44
0.9	-1.93	-2.01	-1.66	-1.68
0.95	-2.51	-2.59	-1.97	-2.03
0.975	-3.21	-3.31	-2.51	-2.62
0.99	-4.14	-4.49	-3.23	-3.07
0.995	-4.72	-5.26	-3.64	-3.21

The impact on future earning was also tested for proportion of poor on a regional level, NUTS2 and NUTS3, see Table 12. In accordance with the above results, the lowest/smallest geographical units have the larger negative effect. That is, we found regional differences on later life income depending on the level of poverty in the region. The effect sizes are small. Coming from a NUTS3 region with a poverty rate one standard deviation above the mean reduces the expected young adult income with 50.4% of a percentile. The corresponding figure for NUTS2 regions is 0.48%.

Table 12: Models with individual and family characteristics and proportion of poor in NUTS2 and NUTS3 respectively explaining income in young adulthood. Controls as in Table 1.

	(1) Income	Std Error	(2) Income	Std Error
NUTS3, mean poor (County)	-42.229***	5.582		
NUTS2, mean poor			-34.461***	6.220
cons	61.769***	.858	60.664***	.941
N	103945		103945	
adj. R^2	.063		.063	

Standard errors in second column. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We also run interactions between neighbourhood level and Non-European background to find out differences in comparison with the European and Swedish background individuals. When interactions are introduced the parameter for non-European background becomes much more negative. This suggests that that we introduce multicollinearity so the preliminary analyses that being of Non-European background increases income must be taken with caution (tables available from authors upon request).

6.3. Conclusion

From the above results we conclude that the local level neighbourhoods are the most influential concerning future income. Also that the poverty level in areas are found to be associated negatively with income. The larger the proportion of poor in the neighbourhood the more negative effect on a person's income in the late 20s.

Another concern resulting from this study is that the variation in individual's income might not have reached its full potential. It might be worth checking for later life variation in income in older cohorts to compare with the variation in income in our cohort. Often neighbourhood effects on education are measured as more influential which might be explained by the earlier completion of education in a person's life course compared to labour-income.

Another issue of concern for further research is that some recent studies in Sweden have shown larger measurable neighbourhood effects from affluence than from poverty. Therefore, a focus on affluent and poor neighbourhoods in comparison could be interesting.

Nevertheless, there is still about the same negative effect on future income from living in a single mother family as from living in an area measuring poor among the 200 closest neighbours. Concerning policies preventing generational transmissions of poverty neighbourhoods are thus worthwhile working with.

7. Results from the longitudinal microdata analysis for the Netherlands

7.1. Introduction

In this section we report the results from our analyses on how contextual poverty at multiple geographical scales is related to individual labour income and obtained educational level later in life on register data from the Netherlands. We also estimated how individual characteristics and family characteristics at age 16 are related to the outcome variables. These individual and family characteristics are also included in all models in which we estimated the effect of contextual poverty as especially family socio-economic characteristics are related to both individual socio-economic outcomes and the residential area were the individual was living at age 16.

In order to better understand the effect of contextual poverty, we measured it at two types of geographies and multiple scales. First we constructed bespoke geographies and calculated the poverty rate at 5 different spatial scales. We calculated the proportion of individuals with a low income of the nearest 200, 1600, 12800, 51200, and 204800 neighbours. In addition, we used more aggregated and fixed boundary administrative units. We calculated the poverty rate at NUTS3 (41 regions) and NUTS2 level (12 regions, equivalent to provinces). Using measures of contextual poverty at multiple geographical scales allows us to test hypotheses on contextual effects on individual incomes at the level of regions all the way down to the very local environment of the nearest 200 neighbours.

Comparing the poverty rate at these different spatial scales indicates that most variation can be seen at the lowest spatial scale (k=200), with the contextual poverty rate ranging from 0 to .90, meaning that in the most affluent area 0% of the nearest 200 individuals has low income, and in the most deprived area 90% of the nearest 200 individuals has a low income. The distribution becomes smaller with increasing spatial scale (see Mean, SD, Min and Max in Table 14). At the highest spatial scale (k=204800), contextual poverty ranges from 6% individuals with a low income in the most affluent area to 21% individuals with a low income in the most deprived area. At low spatial scales populations are more homogeneous than in larger regions, hence variation in measures of poverty concentrations at different scales.

Table 13: Descriptive statistics of individual and family characteristics

Variable	Mean	Std. Dev.	Min	Max
Individual characteristics				
Income age 30 (percentiles)	69.99	22.36	1	100
Male	0.51		0	1
Non-European migration background	0.14		0	1
Family characteristics				
Single parent family when 16 years	0.14		0	1
Social allowance when 16 years	0.04	0.19	0	1
Tertiary education parents				
One parent	0.15		0	1
Two parents	0.03		0	1
missing	0.39		0	1
Household income when 16 years (thousand Euros)	20.82	13.73	-[a]	-[a]
Unemployment parents when 16 years				
One parent	0.27		0	1
Two parents	0.04		0	1
Contextual characteristics				
Poverty rate *k*=200	0.09	0.06	0	0.90
Poverty rate *k*=1600	0.09	0.04	0.01	0.41
Poverty rate *k*=12800	0.09	0.03	0.03	0.33
Poverty rate *k*=51200	0.10	0.03	0.05	0.26
Poverty rate *k*=204800	0.10	0.02	0.06	0.21
Poverty rate NUTS3	0.10	0.02	0.07	0.13
Poverty rate NUTS2	0.10	0.01	0.08	0.12

Note: [a] Minimum and Maximum values are not reported due to disclosure rules of Statistics Netherlands; $N_{individual}$=158,561; $N_{100 \text{ by } 100 \text{ meter square}}$ = 111,184; N_{NUTS3}=41; N_{NUTS2}=12.
Source: System of Social statistical Datasets(SSB).

7.2. Discussion of main findings

First we present the results of the models predicting individual earnings at age 30. Table 15 presents results of a model in which only individual and family characteristics are included. The results for the two individual characteristics show that males earn significantly more than females and that individuals with a non-European migration background earn less compared to individuals from a European background. The difference in income percentile between individuals with and without a non-European background is on average 4.8 percentiles.

We also included a range of variables that present the socio-economic status of the family when the individual was 16 years old. Individuals who had a single parent at age 16 have a lower income compared to individuals who had a two-parent family; the difference is 3.7

percentiles. Having one or two parents with tertiary education is related to a higher income compared to having no parents with tertiary education. The difference between only one or both parents having obtained tertiary education, however, does not make a large difference. Household income when the individual was 16 years old is positively related to income, meaning that a higher household income at age 16 is related to a higher individual income at age 30. For every ten thousand Euros family income was higher, the income of an individual ended up 1.3 percentiles higher compared to other individuals. Parental unemployment at age 16 is also strongly related to individual income ate age 30. Having one unemployed parent is related to an income 2.4 percentiles lower compared to no unemployed parents. Two unemployed parents is even more disadvantageous. The difference in income percentile between an individual who had two unemployed parents when he or she was 16 and an individual who did not have an unemployed parent is as large as 5.2 percentiles. In total these individual and family characteristics explained 8.6% of the variance in individual income.

Table 14: Individual earnings at age 30 (percentile) predicted by individual and family characteristics

	B	SE
Male	9.380***	.107
Non-European migration background	-4.820***	.165
Single parent family when 16 years	-3.661***	.164
Social allowance when 16 years	-2.841***	.336
Tertiary education parents		.
One parent	3.775***	.162
Two parents	4.196***	.315
missing	2.983***	.122
Household income when 16 years (thousand Euros)	.128***	.004
Unemployment parents when 16 years		.
One parent	-2.413***	.126
Two parents	-5.169***	.309
Constant	62.794***	.141
Adjusted R^2	.086	

Note: N= 158,561; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
Source: System of Social statistical Datasets(SSB).

In Table 3 we present the results from the models that estimated the relation between contextual poverty at age 16 and individual income at age 30. At first sight the effect sizes indicate that contextual poverty has a stronger effect on individual income with increasing scale. In order to get a better picture of this relationship at different geographical scales, the distribution of contextual poverty at each spatial scale (see Section **¡Error! No se encuentra el origen de la referencia.**) needs to be taken into account. Using the effect sizes reported in

Table 4 to calculate differences in individual income between individuals from the most affluent areas and the most deprived areas we can say the following. At the lowest spatial scale (k=200), the difference in income at age 30 between an individual who lived in the poorest neighbourhood at age 16 and an individual who lived in the richest neighbourhood at age 16 is 22.3 percentiles. This is the differences between someone from an area with a poverty rate of 0% and someone from an area with a poverty rate of 90%. At moderate spatial scale (k=12800) the differences in individual income between the poorest and richest area is 14 percentiles, which is the differences between an individual from an area with a poverty rate of 3% and an individual from an area with a poverty rate of 33%. At the highest spatial scale (k=204800) this difference is 7.3 percentiles. The richest area at this spatial scale has a poverty rate of 6% and the poorest area a poverty rate of 21%.

The effects of the poverty rate at NUTS3 and NUTS2 levels and individual income are presented in Table 5. In these models as well, all individual and family characteristics were included. Both NUTS-levels are large administrative geographical units. The poverty rate of such large scales show little variation and are close to the national average poverty rate. The most deprived NUTS3 region has a poverty rate of 13%, whereas the most affluent area has a poverty rate of 7%. The difference in income at age 30 between an individual from the poorest and an individual from the most affluent NUTS3 region is 5.7 percentile. At NUTS2-level, which represents the 12 provinces of the Netherlands, the lowest poverty rate is 8% and the highest 12%. The difference in income between an individual from the poorest province and an individual from the most affluent province is 5.4 percentile.

Table 15: Individual earnings at age 30 (percentiles) predicted by contextual poverty (ratios) at age 16 at multiple geographical scales

	k=200		*k*=1600		*k*=12800		K=51200		*k*=204800	
	B	SE	B	SE	B	SE	B	SE	B	SE
Poverty rate (ratio)	-24.768***	.922	-36.787***	1.208	-46.734***	1.544	-51.634***	1.805	-48.450***	2.226
Constant	64.955***	.162	66.058***	.177	67.075***	.200	67.650***	.221	67.427***	.255
Adjusted R^2	.090		.091		.091		.091		.089	

Note: N= 158,561; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Individual and family characteristics included in all models.
Source: System of Social statistical Datasets(SSB).

Table 16: Individual earnings at age 30 (percentiles) predicted by contextual poverty (ratios) at age 16 at NUTS3 and NUTS2 level

	NUTS3		NUTS2	
	B	SE	B	SE
Poverty rate (ratio)	-113.359***	4.242	-135.086***	4.937
Constant	73.618***	.429	75.728***	.493
Adjusted R^2	.090		.091	

Note: N= 158,561; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Individual and family characteristics included in all models.
Source: System of Social statistical Datasets(SSB).

7.3. Conclusion

The results from our analyses on how contextual poverty is related to individual labour income later in life have shown that this relationship differs depending on the geographical scale at which contextual poverty is measured. The difference in income between an individual from the most deprived area and an individual from the most affluent area are most pronounced at low spatial scale. At the lowest spatial scales the population is more homogenous resulting in more extremes when it comes to the proportion of individuals with a low income. Living in a poor neighbourhood at age 16 was related to lower income in the late 20s, net of family socio-economic characteristics. Although family socio-economic characteristics were strongly related to both individual earnings and educational attainment, and are also strongly related to where the individual was living, contextual poverty had an additional effect. At very large spatial scale, NUTS3 (41 regions in the Netherlands) and NUTS2 level (12 provinces of the Netherlands), we found that the poverty rate of these large regions was negatively related to individual income. The problem with interpreting the effects of contextual poverty at such large scales is that the effect of contextual poverty becomes a proxy for a lot of characteristics of the larger areas.

The mechanisms through which the residential or environmental context affects individual outcomes may be different at different spatial scales. The mechanisms range from regional labour markets at larger spatial scales to social network and peer group effects within the immediate environment as captured by the smaller spatial scales. The results have shown that in order to come to a better understanding of the consequences of spatial inequality for individual socio-economic outcomes, it is important to look at spatial inequalities at different geographical scales.

8. Report conclusions

Inequalities among individuals in achieved socio-economic status (e.g. income, employment, education) result both from differences in individual and family characteristics, and differences in contextual characteristics. The results from the analyses using longitudinal microdata for Sweden, the Netherlands and, to less extent, the UK on how contextual poverty, measured in terms of income deprivation, is related to individual labour income have shown that this relationship differs depending on the geographical scale at which contextual poverty is measured. The effect of contextual income deprivation appears to be most pronounced for lower spatial scales. Scaling up to larger geographical areas, the concentrations at micro scale are averaged out, resulting in less extremes of poverty concentration at these scales.

The results have shown that in order to come to a better understanding of the consequences of spatial inequality for individual socio-economic outcomes, it is important to look at spatial inequalities at different geographical scales. However, the approach can only be applied when geocoded data are available for very small spatial units and such data are still unavailable in many countries. Consequently, one very important conclusion and recommendation from the work carried in Work Package 5 of RELOCAL is the need to improve both availability and access to socio-economic geocoded data at very low scale for more countries. Without such type of information, it is very difficult to attempt to provide guidance to policy makers on the more appropriate scales for public policy intervention.

Finally, while our multi-scale approach shows that inequality is a multi-scale problem, on its own it cannot explain which mechanisms operate at different levels. Understanding our modelling outcomes for the different scales requires combining them with detailed case study analysis.

References

ALVES, N. 2012. A view on income redistribution in Portugal and in the European Union. *Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies*.

ALVES, N., MARTINS, C. J. E. B. & ARTICLES, F. S. R. 2012. Mobility and income inequality in the European Union and in Portugal.

ANDERSSON, E. K. & MALMBERG, B. 2016. Segregation and the effects of adolescent residential context on poverty risks and early income career: A study of the Swedish 1980 cohort. *Urban Studies*, 0042098016643915.

ANDERSSON, R., MUSTERD, S., GALSTER, G. & KAUPPINEN, T. M. 2007. What Mix Matters? Exploring the Relationships between Individuals' Incomes and Different Measures of their Neighbourhood Context. *Housing Studies,* 22, 637-660.

ARISTEI, D. & PERUGINI, C. J. E. S. 2015. The drivers of income mobility in Europe. 39, 197-224.

ATKINSON, A. B., GUIO, A.-C. & MARLIER, E. 2017. *Monitoring social inclusion in Europe*, Publications Office of the European Union.

BAKKER, B. F. M. 2002. *Statistics Netherlands' approach to social statistics: The Social Statistical Dataset*, OECD Statistics Newsletter.

BAUER, T. K., FERTIG, M. & VORELL, M. 2011. The Effect of Neighborhood Characteristics on Individual Employment Probability. *Ruhr Economic Papers,* 285.

BERGER, M., SCHAFFNER, S. J. J. O. E. & MEASUREMENT, S. 2016. A note on how to realize the full potential of the EU-SILC data. 41, 395-416.

BOLSTER, A., BURGESS, S., JOHNSTON, R., JONES, K., PROPPER, C. & SARKER, R. 2007. Neighbourhoods, households and income dynamics: a semi-parametric investigation of neighbourhood effects. *Journal of Economic Geography,* 7, 1-38.

BORST, M. 2018. *EU-SILC Tools: eusilcpanel - first computational steps towards a cumulative sample based on the EU-SILC longitudinal datasets,* DEU, Köln.

BRÄNNSTRÖM, L. 2005. Does Neighbourhood Origin Matter? A Longitudinal Multilevel Assessment of Neighbourhood Effects on Income and Receipt of Social Assistance in a Stockholm Birth Cohort. *Housing, Theory and Society* 22, 169-195.

BUCK, N. 2001. Identifying neighbourhood effects on social exclusion. *Urban Studies,* 38, 2251-2275.

CHAMBERLAIN, G. 1982. Multivariate Regression Models for Panel Data. *Journal of Econometrics,* 1, 5-46.

CHESHIRE, P. C. 2007. Segregated neighbourhoods and mixed communities: a critical analysis. York, UK: Joseph Rowntree Foundation.

CLARK, W. A. V., VAN HAM, M. & COULTER, R. 2014. Spatial mobility and social outcomes. *Journal of Housing and the Built Environment,* 29, 699-727.

Co-funded by the Horizon 2020
Framework Programme of the European Union

DIETZ, R. D. 2002. The estimation of neighborhood effects in the social sciences: an interdisciplinary approach. *Social Science Research,* 31**,** 539–575.

DUJARDIN, C. & GOFFETTE-NAGOT, F. 2007. Neighborhood effects, public housing, and unemployment in France (Working paper 05-05). *Paris, France: GATE*.

DUJARDIN, C. & GOFFETTE-NAGOT, F. 2010. Neighborhood effects on unemployment?: A test à la altonji. *Regional Science and Urban Economics,* 40**,** 380-396.

DUJARDIN, C., PEETERS, D. & THOMAS, I. 2009. Neighbourhood effects and endogeneity issues. *Université catholique de Louvain . Center for Operations Research and Econometrics (CORE), CORE Discussion*

DURLAUF, S. N. 2004. Chapter 50 - Neighborhood Effects. *In:* HENDERSON, J. V. & JACQUES-FRANÇOIS, T. (eds.) *Handbook of Regional and Urban Economics.* Elsevier.

ELLEN, I. G. & TURNER, M. A. 1997. Does neighborhood matter? Assessing recent evidence. *Housing Policy Debate,* 8**,** 833–866.

ENGEL, M. & SCHAFFNER, S. 2012. *How to use the EU-SILC panel to analyse monthly and hourly wages*, Ruhr Economic Papers.

EUROFOUND 2009.

FERREIRA, F. H. G. & GIGNOUX, J. 2011. THE MEASUREMENT OF INEQUALITY OF OPPORTUNITY: THEORY AND AN APPLICATION TO LATIN AMERICA. 57**,** 622-657.

FERREIRA, F. H. G. & GIGNOUX, J. 2014. The Measurement of Educational Inequality: Achievement and Opportunity1. *World Bank Economic Review,* 28**,** 210-246.

GALSTER, G. 2002. An economic efficiency analysis of deconcentrating poverty populations. *Journal of Housing Economics,* 11**,** 303–329.

GALSTER, G., ANDERSSON, R., MUSTERD, S. & KAUPPINEN, T. M. 2008. Does neighborhood income mix affect earnings of adults? New evidence from Sweden. *Journal of Urban Economics* 63**,** 858–870.

GEOSTAR 2015. Geographical context; New ways of measuring the significance of surrounding for the life course of individuals (Geografisk kontext: Ett nytt sätt att mäta vad omgivningen betyder för individens livsbana). MONA (Micro Data on-line Access) Statistics Sweden accessed via Department of Human Geography, Stockholm University: Statistics Sweden.

HEDMAN, L. & GALSTER, G. 2013. Neighbourhood income sorting and the effects of neighbourhood income mix on income: A holistic empirical exploration. *Urban Studies,* 50**,** 107-127.

HEDMAN, L., HAM, M. V. & MANLEY, D. 2011. Neighbourhood Choice and Neighbourhood Reproduction. *Environment and Planning A: Economy and Space,* 43**,** 1381-1399.

HEDMAN, L., MANLEY, D., HAMAND, M. V. & OSTH, J. 2015. Cumulative exposure to disadvantage and the intergenerational transmission of neighbourhood effects. *Journal of Economic Geography* 15**,** 195–215.

Co-funded by the Horizon 2020
Framework Programme of the European Union

HENNERDAL, P. 2019. *geocontext, GitHub repository*. *In:* NEIGHBOURS), A. P. S. T. A. T. K.-N. (ed.). https://github.com/PonHen/geocontext.

HOUBIERS, M. 2004. Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of official statistics,* 20**,** 55.

IOANNIDES, Y. M. & ZABEL, J. E. 2008. Interactions, neighborhood selection and housing demand. *Journal of Urban Economics,* 63**,** 229–252.

JANSSEN, H. J., VAN HAM, M. & (WITH CONTRIBUTIONS FROM MELO P., A. E., MALMBERG B., FRITSCH M. & NÉMETH S.) 2018. D5.2 Report on multi-scalar patterns of inequalities. H2020 project RELOCAL – Resituating the Local in Cohesion and Territorial Development.

JUÁREZ, F. W. C. & SOLOAGA, I. 2014. iop: Estimating ex-ante inequality of opportunity. *The Stata Journal,* 14**,** 830-846.

MCCULLOCH, A. 2001. Ward-level deprivation and individual social and economic outcomes in the British household panel study. *Environment and Planning A,* 33**,** 667-684.

MELLANDER, C., STOLARICK, K. & LOBO, J. 2017. Distinguishing neighbourhood and workplace network effects on individual income: evidence from Sweden. *Regional Studies,* 51**,** 1652-1664.

MUNDLAK, Y. 1978. On the pooling of time series and cross sectional data. *Econometrica,* 46**,** 69–85.

NIELSEN, M. M., HAANDRIKMAN, K., CHRISTIANSEN, H., COSTA, R., SLEUTJES, B., ROGNE, A. & STONAWSKI, M. 2017. Residential Segregation in 5 European Countries. Retrieved from http://docs.wixstatic.com/ugd/870ecc_d148b555abb542d19bfcb0c1358e0f17.pdf.

OECD 2015. In It Together: Why Less Inequality Benefits All.

ÖSTH, J., MALMBERG, B. & ANDERSSON, E. 2014. Analysing segregation with individualized neighbourhoods defined by population size. *In:* LLOYD, C. D., SHUTTLEWORTH, I. & WONG, D. (eds.) *Social-Spatial Segregation: Concepts, Processes and Outcomes.* Policy Press.

PLOTNICK, R. D. & HOFFMAN, S. D. 1999. The effect of neighborhood characteristics on young adult outcomes: alternative estimates. *Social Science Quarterly,* 80**,** 1-18.

POLIN, V. & RAITANO, M. 2014. Poverty Transitions and Trigger Events across EU Groups of Countries: Evidence from EU-SILC. *Journal of Social Policy,* 43**,** 745-772.

SARI, F. 2012. Analysis of neighbourhood effects and work behaviour: Evidence from Paris. *Housing Studies,* 27**,** 45–76.

SHORROCKS, A. 1978. MEASUREMENT OF MOBILITY. *Econometrica,* 46**,** 1013-1024.

UNIVERSITY OF ESSEX. INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH, N. S. R., KANTAR PUBLIC. 2016. Understanding Society: Waves 1-6, 2009-2015: Special Licence Access. 7th Edition. UK Data Service. SN: 6931.

VAN HAM, M., BOSCHMAN, S. & VOGEL, M. 2018. Incorporating Neighborhood Choice in a Model of Neighborhood Effects on Income. *Demography,* 55**,** 1069-1090.

VAN HAM, M. & MANLEY, D. 2010. The effect of neighbourhood housing tenure mix on labour market outcomes: a longitudinal investigation of neighbourhood effects. *Journal of Economic Geography,* 10**,** 257–282.

VAN HAM, M. & MANLEY, D. 2012. Neighbourhood Effects Research at a Crossroads. Ten Challenges for Future Research Introduction. *Environment and Planning A,* 44**,** 2787-2793.

VAN HAM, M. & MANLEY, D. 2015. Occupational Mobility and Living in Deprived Neighbourhoods: Housing Tenure Differences in 'Neighbourhood Effects'. *Applied Spatial Analysis and Policy,* 8**,** 309-324.

VAN KERM, P. & ALPERIN, M. 2013. Inequality, growth and mobility: The intertemporal distribution of income in European countries 2003-2007. *Economic Modelling,* 35**,** 931-939.

WIMARK, T., HAANDRIKMAN, K. & NIELSEN, M. M. 2019. Migrant labor market integration: The association between initial settlement and subsequent employment and income among migrants. *Geografiska Annaler. Series B, Human Geography*.

Appendix A

Overview of modelling approaches to address identification issues

Sample restriction

This is the simplest method and consists of constraining the study sample to a sub-set of individuals for whom residential location decision can be assumed to be exogenous. Typically, this implies restricting the sample to young adults (e.g. aged 19 to 25) residing with at least one parent (e.g. Dujardin et al., 2009). The assumption is that the choice of a residential location has been made previously by the parents and is thus exogenous to the employment and/or income status of their children. There are some limitations: first, the sample restriction may not completely eliminate endogeneity because it is likely that parental characteristics determining residential choice may also influence children's future well-being outcomes; second, the method can itself generate selection bias since young adults living with their parents may not be representative of the population of young adults, and can actually be more likely to include individuals who are less able to afford living on their own: and finally, this method results in a large reduction in sample size.

In the attempt to rule out residential sorting bias by design, some researchers used data for individuals up to the point they move residential location and/or those individual who never moved residence. This is, constraining the sample to non-movers for whom residential sorting would not apply or apply less strongly (e.g. Buck, 2001).

Other sample restriction approaches include constraining the sample to siblings by focusing on children belonging to the same family (and hence supposedly exposed to the same family contextual factors) but who have grown up in different neighbourhoods because of family relocation (e.g. Plotnick and Hoffman, 1999). One obvious limitation is data availability since survey data studies are not likely to have a large enough sample to make this method feasible. By combining siblings' data with family fixed-effects, models would be able to capture in a cleaner way the effect from childhood neighbourhood on children's outcomes as adults. However, the assumption that household residential preferences do not change over time, and more specifically with the number (and possibly gender) of children, is questionable. As a result, family relocation may reveal a change in unobservable household preferences and

Co-funded by the Horizon 2020
Framework Programme of the European Union

concerns with children's local environment (e.g. crime, school quality), which in turn are likely to affect children's outcome. Therefore, this method may not fully eliminate endogeneity because it is likely that parental characteristics determining residential choice may also influence children's future well-being outcomes.

A less common approach was used by (Clark et al., 2014), which consisted of studying only one randomly selected individual from each household in the BHPS at its initial year instead of using all adults in each household. By randomly selecting one individual in each household at the beginning of the survey life (i.e. 1991 and in 1999 for the Welsh and Scottish booster samples) and tracking only those randomly selected individuals over time, the authors claimed to reduce the potential for residential selection bias because keeping all adult household members in the same household could bias the analysis against relocation decisions of smaller households. This approach results in a considerable reduction of sample size without safely removing residential sorting bias given that we don't know how much each individual in a household influences relocation decisions in what is generally a household-level decision involving negotiations between the various members of the same household. Moreover, it is not clear how this approach deals with individuals who move into and out of different types of households (e.g. divorce, marriage).

Longitudinal data and individual fixed-effects

With increasing availability of survey-based longitudinal microdata, some studies have attempted to account for selection bias by using estimators based on individual fixed-effects. This method relies on strong assumptions, in particular that sorting operates on the basis of time-invariant unobservable heterogeneity only (e.g. innate ability, taste). However, if one suspects that time-variant factors may also influence the choice of location, for example changes in preferences over the life cycle, the individual fixed-effects correction will not be sufficient.

There are important downsides to using a fixed-effects estimator. The main disadvantage is that it does not allow estimating the coefficient of time-invariant regressors (e.g. educational attainment, ethnicity) and only considers within-individual variation. Moreover, because it

Co-funded by the Horizon 2020
Framework Programme of the European Union

uses only within-individual variation, it can result in a great loss of efficiency in the estimation of effects for variables that tend to be time persistent (i.e. that change slowly over time).

An alternative and more flexible approach to the fixed-effects model is to consider a hybrid approach combining the fixed-effects with the random-effects approach, the correlated random-effects model (Mundlak, 1978, Chamberlain, 1982). The correlated random-effects approach permits estimating the effects of time-invariant covariates (e.g. ethnicity, educational attainment) while still allowing for correlation between individual heterogeneity and the model’s covariates. As an example, Hedman et al. (2015) combine individual fixed- and random-effects estimators in their study of neighbourhood poverty on personal income for Stockholm, Sweden.

Instrumental variables

The use of instrumental variables (IV) techniques to correct for endogeneity issues requires finding variables (i.e. instruments) that influence the choice of residential location in deprived or poor neighbourhoods, but do not affect individual outcome other than through the endogenous neighbourhood-level variables. Finding good instruments is generally a very difficult task. Popular instruments for neighbourhood selection, particularly selection into deprived neighbourhoods, include the gender mix of household's children (e.g. Dujardin and Goffette-Nagot, 2010). This variable is thought to be relevant because it correlates well with some of the factors influencing allocation of households to social housing, particularly household size, and the fact there is a positive correlation between the presence of social housing in a given area and that area’s level of deprivation. In some countries, such as France, household size and in particular the number of children are part of the criteria used for allocating of households to social housing. The rationale for using gender mix of household's children as an instrument for household size and thus the likelihood to be allocated to social housing is that: i) larger families get preferential access to social housing and are less likely to leave it because of the difficulty in finding other affordable housing if their income is low; and ii) parents with mix-gender children are less likely to have a third or fourth child, which is accompanied by a lower probability of being allocated social housing.

Some of the studies using children gender mix as instruments, also use instruments based on a dummy variable for the spouse's workplace being located in parts of urban areas where deprivation tends to be higher: Dujardin and Goffette-Nagot (2007, 2010) and Dujardin et al. (2009). Other instruments used to control for individual unobservable effects include variables that affect individual income, but not neighbourhood income mix. Examples include categorical variables for: whether an individual was on sick leave during the year; on parental leave during the year; and/or receiving preretirement benefits. Similarly, to control for possible neighbourhood unobservable effects, instruments have been defined based on variables that affect neighbourhood income mix in a given year but not the individual’s income earned during that year. Such instruments include, for example, the individual-partner ethnic combination; individual-partner ethnic combination interacted by number of children; individual-partner ethnic combination interacted by partner income; and the proportion of male children in the household. Examples of studies using these types of instruments include Hedman et al. (2013). The IV method is sometimes combined with other methods, namely the use of individual fixed-effects models (e.g. Hedman and Galster, 2013), the estimation of a system of equations (e.g. Dujardin and Goffette-Nagot, 2007, Dujardin and Goffette-Nagot, 2010), or a control function approach (e.g. Bauer et al., 2011).

Control function based on hedonic house prices

This method has been rarely used in this literature, possibly due to data limitations preventing it to be more widely used. Bauer et al. (2011) combine this method with an IV approach to estimate the effect of the neighbourhood unemployment rate on the probability of individual unemployment. Besides using the IV method to control for correlation between individual unobservable characteristics and neighbourhood characteristics, they estimate a control function based on a hedonic house price model to correct for potential bias of the endogenous neighbourhood effect (i.e. neighbourhood unemployment rate) which may arise through correlation between unobserved and observed neighbourhood characteristics. The control function works by taking the average regional residual from a hedonic house price regression, which includes neighbourhood characteristics and dwelling-specific characteristics. The average residual calculated over each location is used as an additional control variable in the main neighbourhood effects model. The assumption made is that the average regional residual should capture all factors influencing the house price besides the observable characteristics of the individual building and neighbourhood. In other words, it can be viewed as controlling for unobservable regional amenities for which individuals have a greater preference and willingness to pay for.

Explicit modelling of neighbourhood selection behaviour

This approach is the most interesting because it explicitly models the process of residential self-selection, rather than just trying to correct for it. By also providing information on the selection process itself, it offers insights on the linkages between residential sorting and area effects. However, this method is very data demanding. Notable examples include Ioannides and Zabel (2008) for Metropolitan Statistical Areas (MSAs) in the US, Hedman et al. (2011) for Uppsala in Sweden, Sari (2012) for Paris in France, and van Ham et al. (2018) for Utrecht in the Netherlands. The approach consists of using a two-step identification strategy to disentangle the selection process in the relationship between neighbourhood deprivation and individual well-being outcomes. The first step comprises the estimation of a discrete choice model of neighbourhood residential selection, from which researchers can compute the conditional probability that an individual will select a specific area over a choice set of alternative areas. These probabilities are used to construct correction components to adjust parameter estimates in the neighbourhood effects model to be estimated in the second step. There are different approaches to the definition of the choice set (e.g. full choice set or partial choice set using randomization) and to the construction of the correction parameters, but the basic idea is the same. The second step consists of estimating the main neighbourhood effects model, which includes the correction components from the selection model, estimated in step one.

Appendix B

Procedure for merging EU-SILC longitudinal datasets across different releases and countries: Finland, France, and Spain

In the context of the 'Programme of Community action to encourage cooperation between Member States to combat social exclusion' and for producing key policy indicators on social cohesion for the follow up of the EU2020 main target on poverty and social inclusion and flagship initiatives in related domains, we use longitudinal data on individual-level changes over time from the EU reference for comparative statistics on income distribution and social exclusion at the European level from the Survey on income and Living conditions provided by Eurostat.[10] The latest update from April 2018 contains up to 12 different releases (from 2005 to 2016), each release covering four years starting from the 2007 release. The 2005 and 2006 releases cover two and years, respectively, because data for our chosen set of countries are only available starting in 2004.

In this technical Appendix, we briefly describe how we build our cumulative longitudinal sample from 2004 to 2016 from EU-SILC data, which follows a rotational design (a sub-sample is rotated from one year to the next and the other sub-samples remain unchanged) and therefore has year “overlaps” across different releases (described further ahead). This boils down to a combination two kinds of procedures: (i) merge of the four different datasets produced within each release (producing a “master file” for that release); and (ii) merge of the different releases to get the full period coverage.

Countries included in the longitudinal EU-SILC analysis

We exclude countries for which no information exists at the regional level (at least NUTS2), by searching for the variables DB040 and the corresponding flag variable DB040_f for Region contained in the Household Register file (described further ahead). If DB040_f=-1 there is no

[10] https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

Co-funded by the Horizon 2020
Framework Programme of the European Union

information at the regional level. If the length of the string of the variable “region” (db040) is lower than 4, then there only exists information at most at the NUTS-1 level. There are only four countries with information at the NUTS2 level in the EU-SILC longitudinal releases: Spain, France and Finland.

EU-SILC longitudinal structure

The EU-SILC follows a rotational design, whereby a release for any given year contains four sub-samples, *rotational groups*, which have been in the survey for one, two, three or four years. Figure 2 illustrates the rotational design.

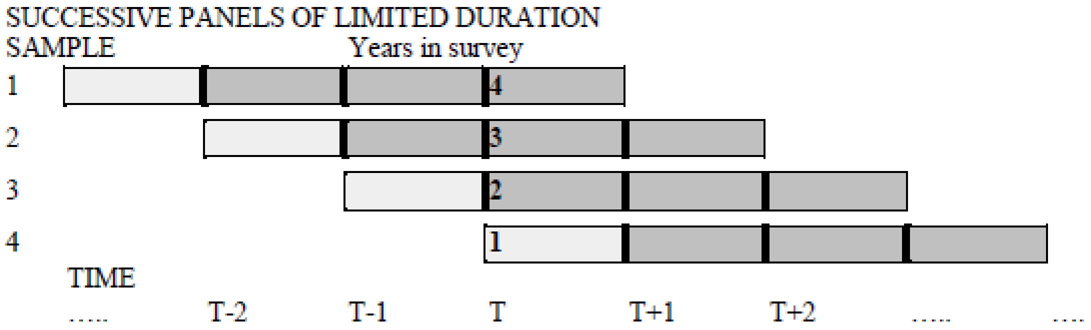


Figure 2: Rotational design.

Source: CIRCABC

Any rotational group remains in the survey for four years (except for the case of France). Each release contains the most recent observations of the rotational groups that are still active, implying that each year one of the four rotational groups from the previous year is lost and a new one is added. Between year t and t+1 the rotational group overlap is around 75%, 50% between t and t+2, 25% between t and t+3 and zero afterwards.[11] For a more detailed analysis on the rotational design followed by EU-SILC and the selection of rotational groups, we refer

[11] This is an approximation. However, in practice, the final dataset has considerable variation in the number of observations over the years for some countries. This is expected for the first and last years but not in between.

the reader to papers such as Atkinson, Guio, and Marlier (2017), Engel and Schaffner (2012) and the detailed guidelines of EU-SILC "Description of target variables: Cross-Sectional and Longitudinal", found at CIRCABC[12].

Data files and description

There are four **longitudinal datasets**:

- Household Register (H-file);
- Household Data (D-file);
- Personal Register (R-file);
- Personal Data (P-file).

The H-file contains data collected at the household level during the surveys, such as income and housing costs. The D-file is a register that contains data that was known prior to the survey, including the household ID, country, region, degree of urbanisation etc. The R-file is a personal register file containing data similar to the H-file but at the individual level (personal ID, country ID, household ID) and with more variables. Finally, the P-file contains personal data collected at the individual level in surveys, such as net and gross income values for employed individuals or self-employed individuals.

The different datasets can be combined together using the following **key variables for merging datasets**:

Year, country, Household ID, Personal ID

In any given year, each household can comprise one or several individuals. As such, the household identifier can be repeated, as many times as the number of individuals it contains, in the personal register data file (R-file). At the same time, while in a given specific year individuals can only be allocated to one unique household, they can change households over

[12] https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp.

Co-funded by the Horizon 2020
Framework Programme of the European Union

time, that is, across years, as a result of a series of different life events (e.g. moving in with a new partner, end of a relationship, getting married, divorcing, becoming a widow, etc.). This means that the variable Household ID is not unique across years, only within a given year. In contrast, the Personal ID never changes, even if the person moves to a different household.

Figure 3 illustrates how the different datasets can be merged together. When merging the household files to the personal register file (R-file), the key variables are the combination of *year*, *country* and *household ID*. When merging the personal data file (P-file) with the personal register file (R-file), or any combination of the latter with the other files, the unique identifier is given by *year*, *country*, *household ID* and *personal ID*.

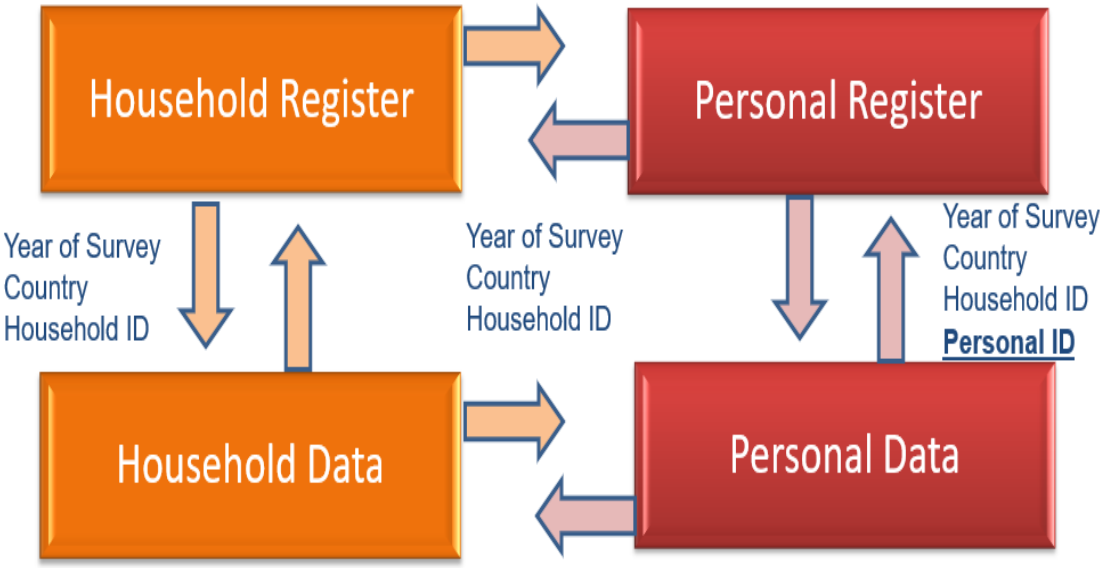


Figure 3: EU-SILC basic structure and linkages

Source: GESIS (Third DwB Training Course, 2014)[13]

[13] http://www.dwbproject.org/events/tc3.html.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Methodology for merging EU-SILC longitudinal datasets

We use STATA 14 to produce the final EU-SILC dataset for the four countries i.e, the combined household-individual-region dataset. We start by merging each of the different files (H, D, R and P) across the releases 2005-2016. This produces what we refer to as master files (master H-file, master D-file, master R-file and master P-file) for each country. After getting these master files, we then merge them together into one final master file according to the design depicted in Figure 4 below.

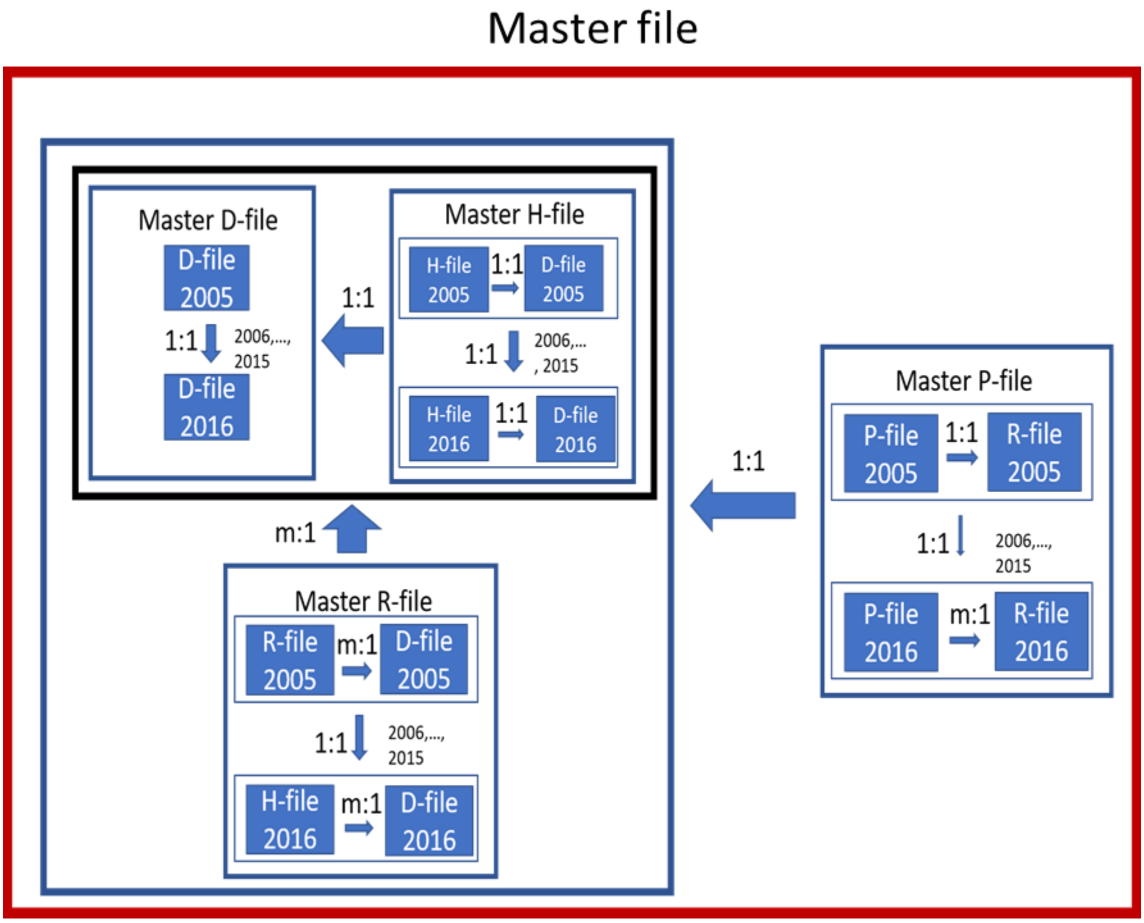


Figure 4: Master file for all releases (authors' own elaboration)

We start with the master household register file (H-file) and combine it with the master household data file (D-file) using, by release, a one-to-one merge over the key variables *year uhid*. We then use the master personal register file (Master R-file) and merge it with the linked H and D files using a many-to-one merge over *year uhid* Finally, we combine the master personal data file (P-file) with the resulting dataset from the previous merges using a one-to-one merge over *year pid*, which renders the complete final master file that includes all available information for the individuals and their respective households.

As reported in Borst (2018), some countries do not strictly follow the EU-SILC rotational design. For our sample, this corresponds to three out of our four countries, namely Finland, Spain and France. We therefore adopt a procedure similar Borst (2018) and use an adaptation of the STATA code produced by GESIS EU-SILC tools (EU-SILCpanel.ado file)[14]. One note of caution is warranted. The GESIS script applies to all releases produced from EUROSTAT up to an update until the release of 2015. More recent updates provided by Eurostat now provide the 4 different files separately by countries. As such, the "EU-SILCpanel.ado" code and should be conveniently adapted to avoid redundancies and thus save on computational effort.

For Finland, Spain and France, there are cases in which the household identifier is not unique when different releases are merged. The code checks whether this is due to a change in the identifier of the rotation group (*rotation*_group) or because the countries re-used the older identifiers. If there is a change in the rotation group when it has already been selected from a more recent release, the group is dropped to avoid overlapping. This means that the rotation group identifiers should range between 1 and 4 across all releases. Therefore, an alternative household ID, *uhid,* is computed that also includes an ID for the rotation group. The latter is called *urtgrp*, which is an alternative ID for rotational groups that is unique across all countries and releases and is a composite string of *country*, *rotation_group*, and the drop-out year of the said group, *drpout_year*. Therefore, *uhid* is also unique across countries and releases. See Borst (2018) for a more detailed description on some of these unique identifiers.

By design, the same individual in EU-SILC can be part of different households in the same year. This means that the combination of *year* and personal ID (*pid*) uniquely identifies entries in

[14] https://www.gesis.org/en/gml/european-microdata/eu-silc/.

Co-funded by the Horizon 2020
Framework Programme of the European Union

personal data file (P file), but not in the personal register file (R file). Therefore, to merge the P file with the R file, a unique personal ID that combines the personal identifier (pid) with the household identifier *hid* and *year* could be computed (as is the case in "eusilcpanel.ado"). However, this alternative unique identifier implies that the same individual can appear repeatedly in the final database, which would cause problems in analysing individual income distributions, measures of inequality of opportunity, and regression analysis. Therefore, we necessarily must drop all duplicates in terms of *year pid* in the R file for each release, before merging it with the combined H+D file.

After compiling the master file, we run the following command in Stata to check whether unique individuals appear in more than 4 years:

```
bysort pid: egen pdcount=count(year)

tab pdcount

tab rotation_group year
```

If *pdcount* takes a value larger than 4, it means that individuals have been interviewed more than 4 years. If the rotation group takes a value greater than 4, it means that the same rotation group ID is being used for different rotation groups across different releases. In either case, these "errors" should not occur, except for the case of France where an individual is followed for 9 years instead of the standard 4 years.[15] [16]

[15] We thank Marwin Borst for helpful insights and comments for building a complete longitudinal EU-SILC dataset.

[16] In Borst (2018), rotational groups in France are followed for 8 years.

Description of final datasets

1. Spain

The master R file for Spain is summarized in the Figure below.

Table 17: Observations in the Master R-file for Spain

YEAR	*Freq.*	*Percent*	*Cum.*
2004	33,851	7.67	7.67
2005	38,271	8.67	16.34
2006	35,307	8.00	24.34
2007	35,250	7.99	32.33
2008	35,879	8.13	40.46
2009	36,547	8.28	48.74
2010	37,066	8.40	57.14
2011	34,818	7.89	65.03
2012	33,484	7.59	72.62
2013	32,247	7.31	79.93
2014	32,131	7.28	87.21
2015	32,966	7.47	94.68
2016	23,470	5.32	100.00
Total	441,287	100.00	

We have a total of 441 287 observations in the master R-file. As for the general master file for Spain, we have 357 220 observations over the period 2004-2016. The variable *pdcount* takes 4 values, ranging between 1-4, as expected.

Table 18: Observations in the master file for Spain

year	Freq.	Percent	Cum.
2004	23,522	6.58	6.58
2005	30,645	8.58	15.16
2006	28,576	8.00	23.16
2007	28,809	8.06	31.23
2008	29,451	8.24	39.47
2009	30,076	8.42	47.89
2010	30,435	8.52	56.41
2011	28,611	8.01	64.42
2012	27,568	7.72	72.14
2013	26,498	7.42	79.56
2014	26,531	7.43	86.98
2015	27,215	7.62	94.60
2016	19,283	5.40	100.00
Total	357,220	100.00	

2. France

The master R-file reports the observations in the table below. We note that the number of observations reported in the master R-file by Borst (2018) is 654 385, almost double our results. What is more, the GESIS report is based on an older EU-SILC update that does not include the 2016 release. We note that we have tested the "eusilcpanel.ado" file for the new update excluding the 2016 release and we get an even lower number of observations (this was also confirmed by Marwin Borst with our adapted code), which indicates reporting errors for France that may pertain to the labelling of rotational groups and its rotational design in general.

Table 19: Observations in the Master R-file for France

year	*Freq.*	*Percent*	*Cum.*
2004	22,141	6.58	6.58
2005	24,458	7.27	13.84
2006	25,269	7.51	21.35
2007	26,227	7.79	29.14
2008	25,813	7.67	36.81
2009	25,955	7.71	44.53
2010	26,836	7.97	52.50
2011	27,369	8.13	60.63
2012	28,949	8.60	69.23
2013	26,758	7.95	77.18
2014	27,263	8.10	85.28
2015	27,073	8.04	93.32
2016	22,468	6.68	100.00
Total	336,579	100.00	

The final master file contains 260 931 individuals. The variable *pdcount* takes 9 values, which is still above the expected 8 (see discussion above). It is possible that a new value for the rotational group was assigned, with individuals that belong to other rotational groups.

We end up with 281 024 observations over the period 2004-2016. However, based on the April 2018 EU-SILC update, the rotational groups seem to, in fact, span over 9 years.

Table 20: Observations in the master file for France

year	Freq.	Percent	Cum.
2004	17,127	6.56	6.56
2005	18,720	7.17	13.74
2006	19,234	7.37	21.11
2007	20,205	7.74	28.85
2008	19,976	7.66	36.51
2009	20,080	7.70	44.20
2010	20,842	7.99	52.19
2011	21,225	8.13	g60.33
2012	22,515	8.63	68.95
2013	20,585	7.89	76.84
2014	21,414	8.21	85.05
2015	21,292	8.16	93.21
2016	17,716	6.79	100.00
Total	260,931	100.00	

3. Finland

The master R-file for Finland reports 270 945 observations, as can be observed in the table below.

Table 21: Observations in the Master R-file for Finland

year	Freq.	Percent	Cum.
2004	15,474	5.71	5.71
2005	19,741	7.29	13.00
2006	18,499	6.83	19.82
2007	17,639	6.51	26.33
2008	17,062	6.30	32.63
2009	16,266	6.00	38.64
2010	19,755	7.29	45.93
2011	23,627	8.72	54.65
2012	22,865	8.44	63.09
2013	26,159	9.65	72.74
2014	27,910	10.30	83.04
2015	27,135	10.01	93.06
2016	18,813	6.94	100.00
Total	270,945	100.00	

The variable *pdcount* spans over a maximum of 4 years (as expected). We end up with 209 625 observations in the final master file for the period 2004-2016, as reported below.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 22: Observations in the final master file for Finland

year	*Freq.*	*Percent*	*Cum.*
2004	12,111	5.78	5.78
2005	14,942	7.13	12.91
2006	14,083	6.72	19.62
2007	13,578	6.48	26.10
2008	13,222	6.31	32.41
2009	12,569	6.00	38.40
2010	15,551	7.42	45.82
2011	18,463	8.81	54.63
2012	17,669	8.43	63.06
2013	20,214	9.64	72.70
2014	21,698	10.35	83.05
2015	21,133	10.08	93.13
2016	14,392	6.87	100.00
Total	209,625	100.00	

Files produced for the releases 2005-2016

- Stata datasets - STATA dta files: ESmaster.dta; FRmaster.dta; FImaster.dta.
- Stata output files - STATA log files: ES.log; FR.log; FI.log.
- Stata script files - STATA do files: ES.do; FR.do; FI.do.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Appendix C

As reported by EUROSTAT (2013 version), the names and codes of NUTS2 regions are given by the following table and we further provide a regional map for NUTS2 regions within each country.

Spain

Table 23:NUTS2 regions in Spain and corresponding codes

Code	NUTS	Version	Name
ES11	NUTS2	2013	Galicia
ES12	NUTS2	2013	Principado de Asturias
ES13	NUTS2	2013	Cantabria
ES21	NUTS2	2013	País Vasco
ES22	NUTS2	2013	Comunidad Foral de Navarra
ES23	NUTS2	2013	La Rioja
ES24	NUTS2	2013	Aragón
ES30	NUTS2	2013	Comunidad de Madrid
ES41	NUTS2	2013	Castilla y León
ES42	NUTS2	2013	Castilla-la Mancha
ES43	NUTS2	2013	Extremadura
ES51	NUTS2	2013	Cataluña
ES52	NUTS2	2013	Comunidad Valenciana
ES53	NUTS2	2013	Illes Balears
ES61	NUTS2	2013	Andalucía
ES62	NUTS2	2013	Región de Murcia
ES63	NUTS2	2013	Ciudad Autónoma de Ceuta (ES)
ES64	NUTS2	2013	Ciudad Autónoma de Melilla (ES)
ES70	NUTS2	2013	Canarias (ES)

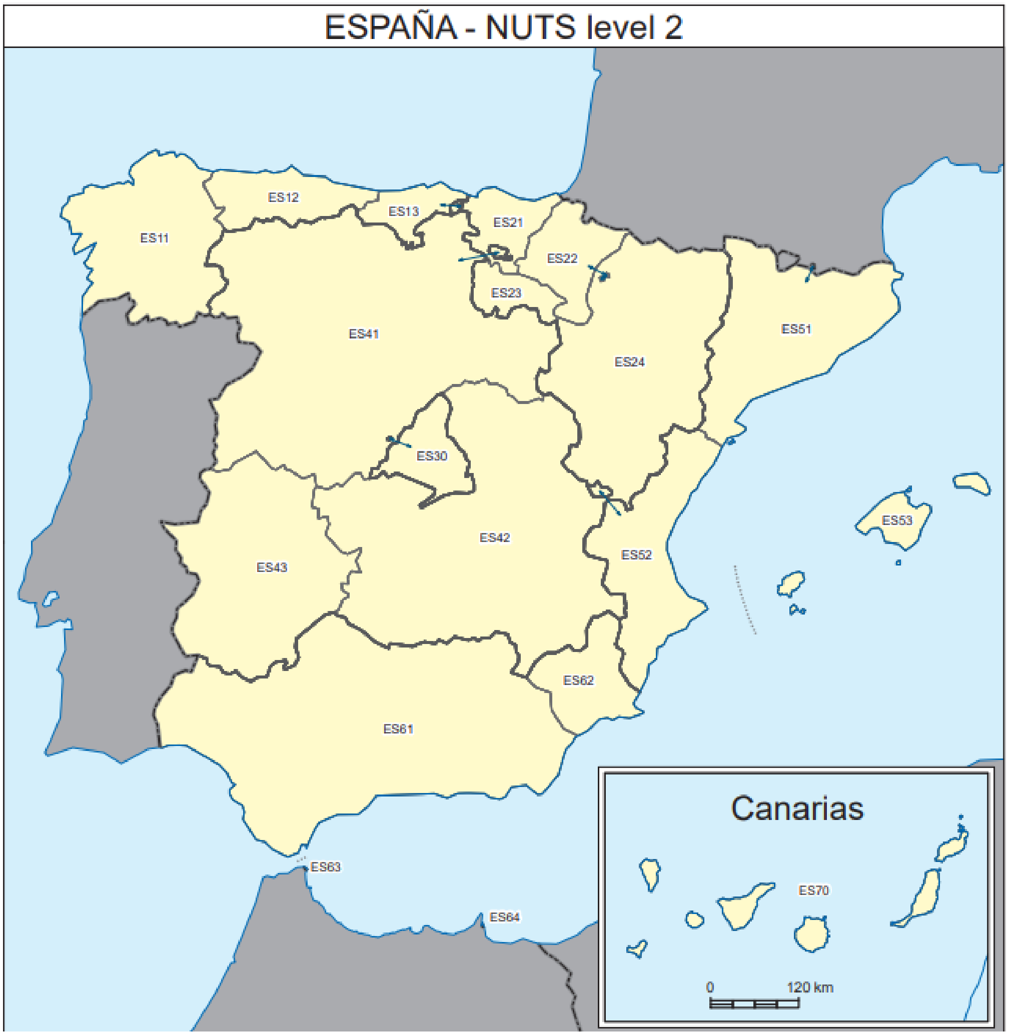


Figure 5: NUTS2 region map for Spain

Source: Eurostat - GISCO, 07/2018

France

Table 24:NUTS2 regions in France and corresponding codes

Code	NUTS	Version	Name
FR10	NUTS2	2013	Île de France
FR21	NUTS2	2013	Champagne-Ardenne
FR22	NUTS2	2013	Picardie
FR23	NUTS2	2013	Haute-Normandie
FR24	NUTS2	2013	Centre (FR)
FR25	NUTS2	2013	Basse-Normandie
FR26	NUTS2	2013	Bourgogne
FR30	NUTS2	2013	Nord - Pas-de-Calais
FR41	NUTS2	2013	Lorraine
FR42	NUTS2	2013	Alsace
FR43	NUTS2	2013	Franche-Comté
FR51	NUTS2	2013	Pays de la Loire
FR52	NUTS2	2013	Bretagne
FR53	NUTS2	2013	Poitou-Charentes
FR61	NUTS2	2013	Aquitaine
FR62	NUTS2	2013	Midi-Pyrénées
FR63	NUTS2	2013	Limousin
FR71	NUTS2	2013	Rhône-Alpes
FR72	NUTS2	2013	Auvergne
FR81	NUTS2	2013	Languedoc-Roussillon
FR82	NUTS2	2013	Provence-Alpes-Côte d'Azur
FR83	NUTS2	2013	Corse
FRA1	NUTS2	2013	Guadeloupe
FRA2	NUTS2	2013	Martinique
FRA3	NUTS2	2013	Guyane
FRA4	NUTS2	2013	La Réunion
FRA5	NUTS2	2013	Mayotte

Co-funded by the Horizon 2020
Framework Programme of the European Union

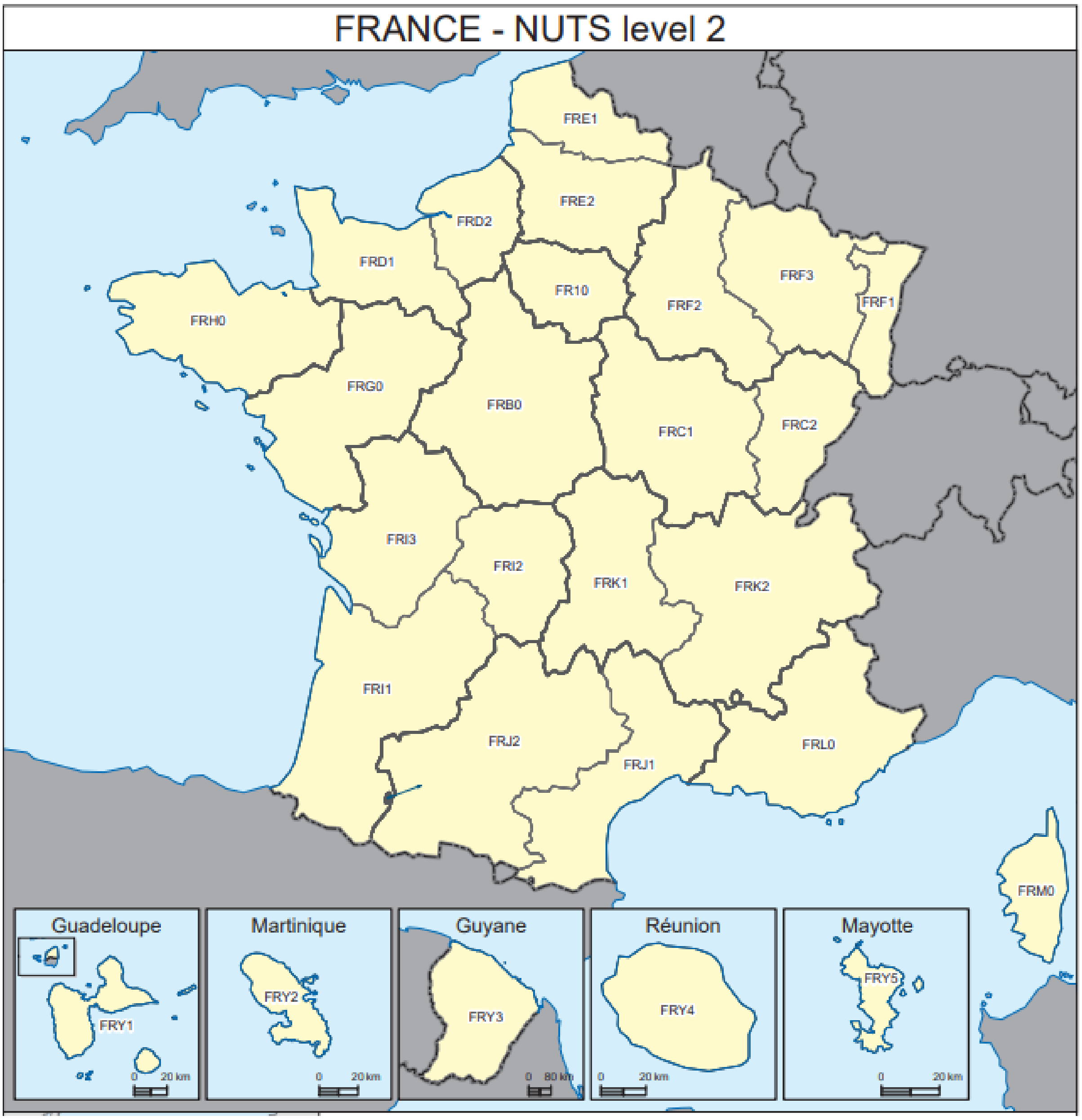


Figure 6: NUTS2 region map for France

Source: Eurostat - GISCO, 07/2018

Co-funded by the Horizon 2020
Framework Programme of the European Union

Finland

Regarding the designation of NUTS2 level regions, they are indexed through a code which we will use always in our output henceforth for the sake of exposition. The codes and corresponding regions, as reported by EUROSTAT (2013 version) are given by the following table, excluding the region of Åland for which no observations were reported.

Table 25:NUTS2 regions in Finland and corresponding codes

Code	NUTS	Version	Name
FI19	NUTS2	2013	Länsi-Suomi
FI1B	NUTS2	2013	Helsinki-Uusimaa
FI1C	NUTS2	2013	Etelä-Suomi
FI1D	NUTS2	2013	Pohjois- ja Itä-Suomi

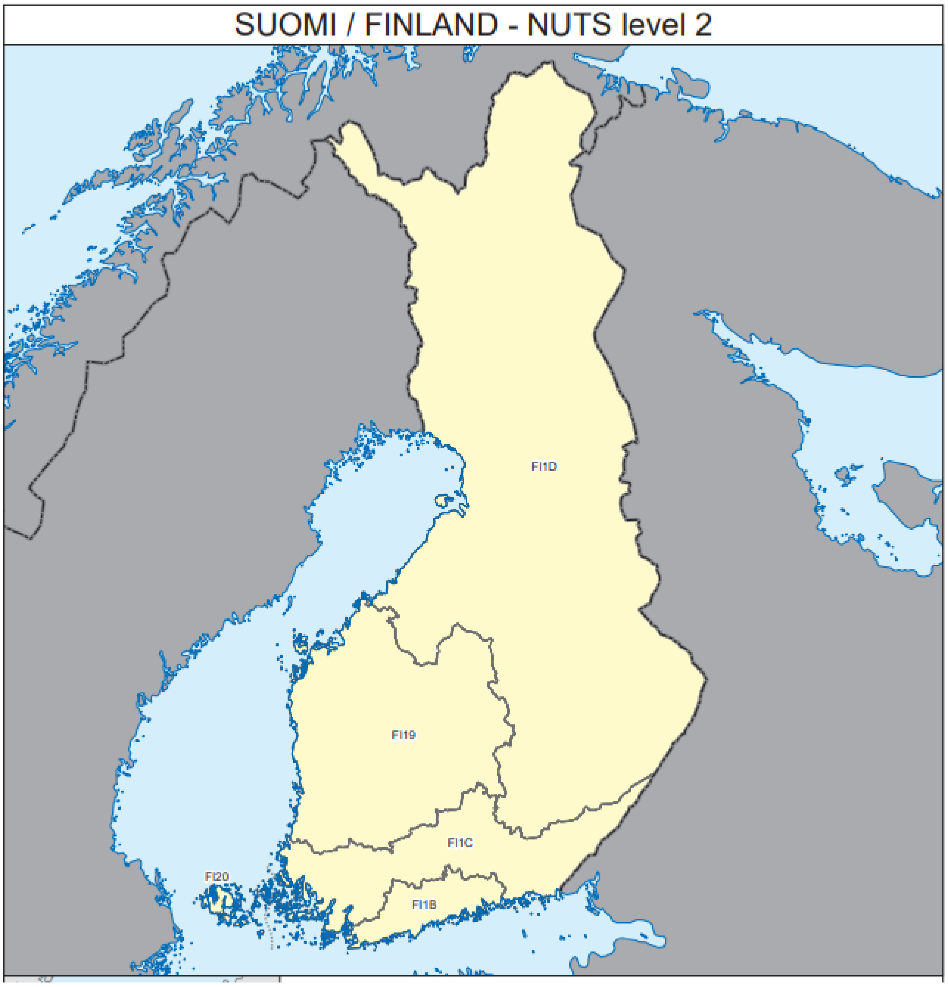


Figure 7: NUTS2 regional map for Finland

Source: Eurostat - GISCO, 07/2018

Co-funded by the Horizon 2020
Framework Programme of the European Union

Appendix D

1. Analysis of income mobility

We report and discuss the results obtained from the analysis of the degree of income mobility at the national level, across NUTS2 regions, and by degree of urbanisation (i.e. large urban areas, small urban areas, and rural areas). Given the four-wave rotational design of EU-SILC, we can only study individuals' income trajectories up to a maximum of four years (if individuals respond to the survey every year). We consider income mobility for 2-year (i.e. transitions between t-1 and t) and 4-year (i.e. transitions between t-3 and t) income trajectories. For the analyses at the three different levels, we report our results here only regarding gross income values. The reason is that the comparisons between gross and net income evidence invariance regarding income mobility. In contrast, when discussing income inequality in the next section, the comparison is warranted as it allows us to infer about the redistributive effects of taxes and social contributions deducted at source.

The concept of income mobility corresponds to the change of a given individual along the income distribution, which we analyze through the construction of transition's position matrices between income deciles. We summarize the results using stacked bar charts showing the percentage of individuals that did not move position in the income distribution, the percentage of individuals that moved to the adjacent decile (i.e., just one decile up and down the income distribution), and the percentage of individuals that moved 2 or more deciles, both up and down the income distribution.

We note that, contrary to many studies who use EU-SILC longitudinal data covering a reduced period (Alves et al., 2012, Alves, 2012, Van Kerm and Alperin, 2013, Polin and Raitano, 2014, Aristei and Perugini, 2015, Berger et al., 2016), our analysis covers a longer period of 11 years. This may contribute to smoother long-run transitions, partially because there is an increased heterogeneity among individuals who participate in the surveys.

Co-funded by the Horizon 2020
Framework Programme of the European Union

- **National level analysis**

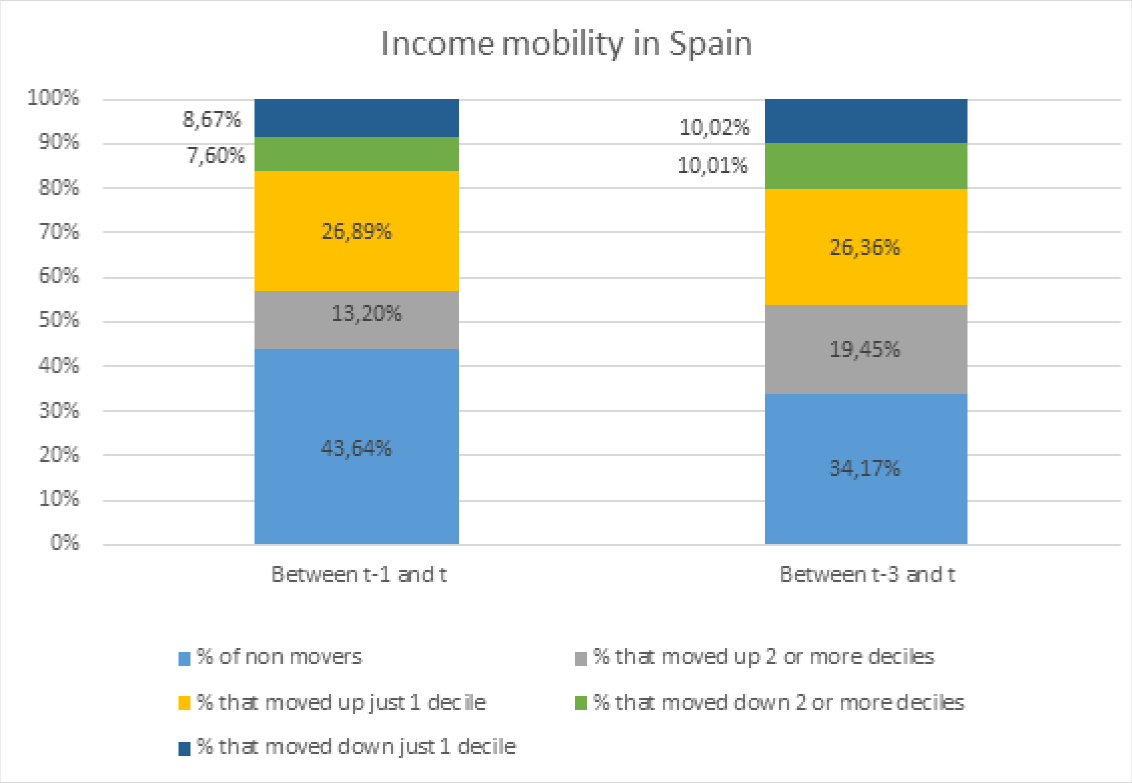


Figure 8: Income mobility in Spain

From Figure 8, we can see that 44% of the individuals remained in the same decile of the income distribution between years t and t-1, whereas 34% did not move along the income distribution between years t and t-3. Of those that moved along the distribution, 40% (16%) of the individuals moved one or more deciles up (down) between t-1 and t, whereas 46% (20%) moved one or more deciles up (down) between t-3 and t. Overall, there is significant income mobility considering both time intervals. Between t-3 and t income mobility was higher, as expected given a higher time span for changes in income. In both cases, upward mobility is mostly restricted to moving up one decile only, corresponding to 27% and 26% of individuals between t-1 and t and t-3 and t respectively. Higher levels of upward mobility (i.e. 2+ deciles) are more achieved over the 4-year period (19%) than over 2 consecutive years (13%).

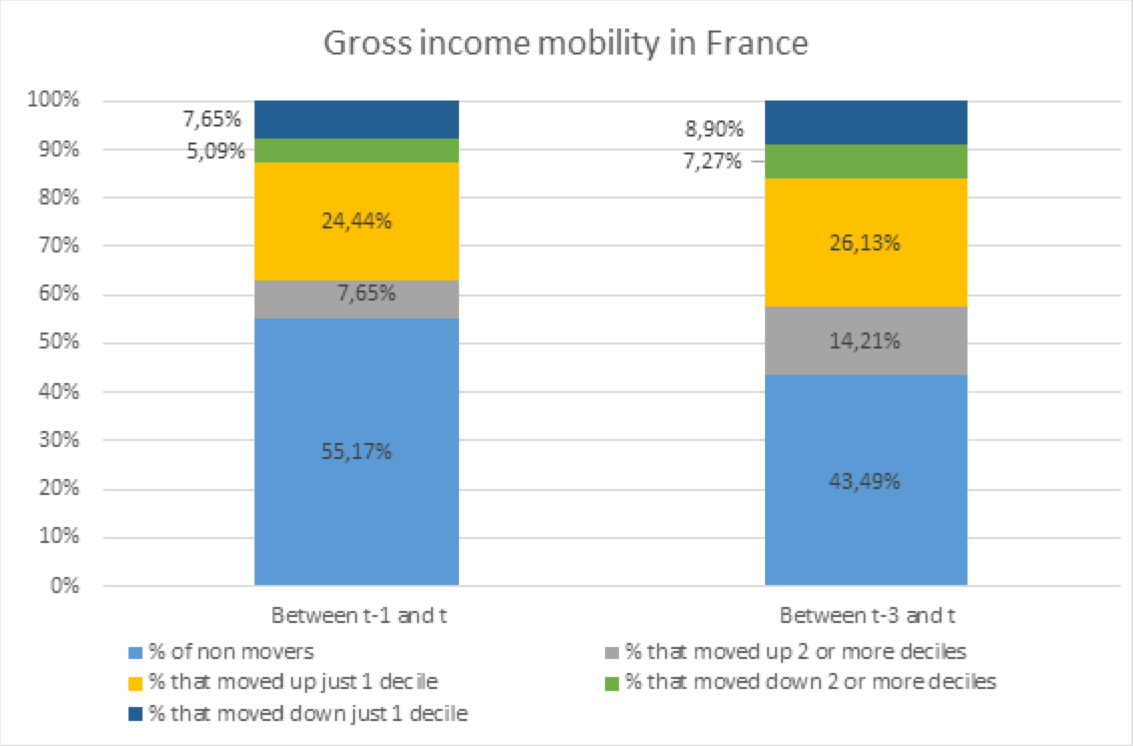


Figure 9: Income mobility in France

The results for France show a very similar picture. By looking at Figure 9, we can see that 55% of the individuals remained in the same deciles of the income distribution between years t and t-1, whereas 43.5% did not move along the income distribution between years t and t-3. Of those that moved along the distribution, about 32% (13%) of the individuals moved one or more deciles up (down) between t-1 and t, whereas 40% (16%) moved one or more deciles up (down) between t-3 and t. Overall, there is significant income mobility considering both time intervals. Between t-3 and t income mobility was higher, as expected given a higher time span for changes in gross income.

Co-funded by the Horizon 2020
Framework Programme of the European Union

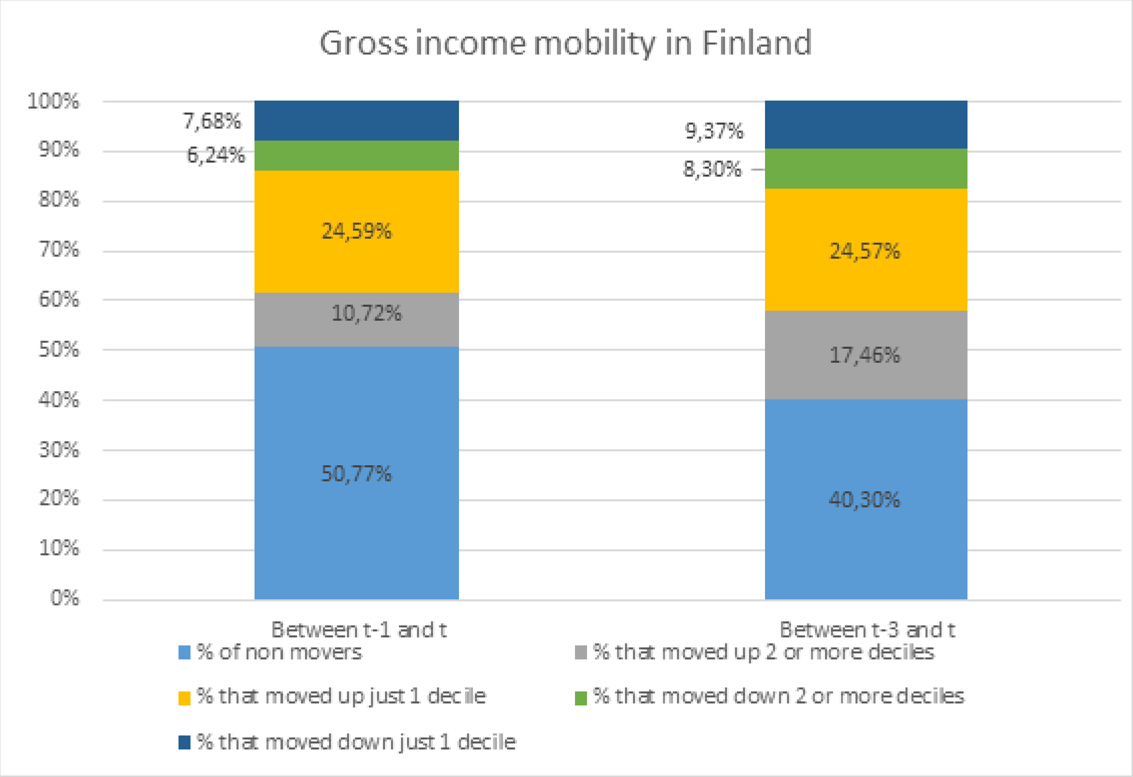


Figure 10: Income mobility in Finland

By looking at Figure 10, we can see that 51% of the individuals remained in the same deciles of the income distribution between years t and t-1, whereas 40% did not move along the income distribution between years t and t-3. Of those that moved along the distribution, about 35% (14%) of the individuals moved one or more deciles up (down) between t-1 and t, whereas about 42% (18%) moved one or more deciles up (down) between t-3 and t. Overall, there is significant income mobility considering both time intervals.

Considering the results for both time periods and the three countries, we can conclude that income mobility was overall higher in Spain, followed by Finland and finally France. This pattern is also true for income upward mobility in both periods.

- **Degree of urbanisation analysis**

We now briefly discuss income mobility by degree of urbanisation of residential areas: large urban areas, small urban areas, and rural areas.

Co-funded by the Horizon 2020
Framework Programme of the European Union

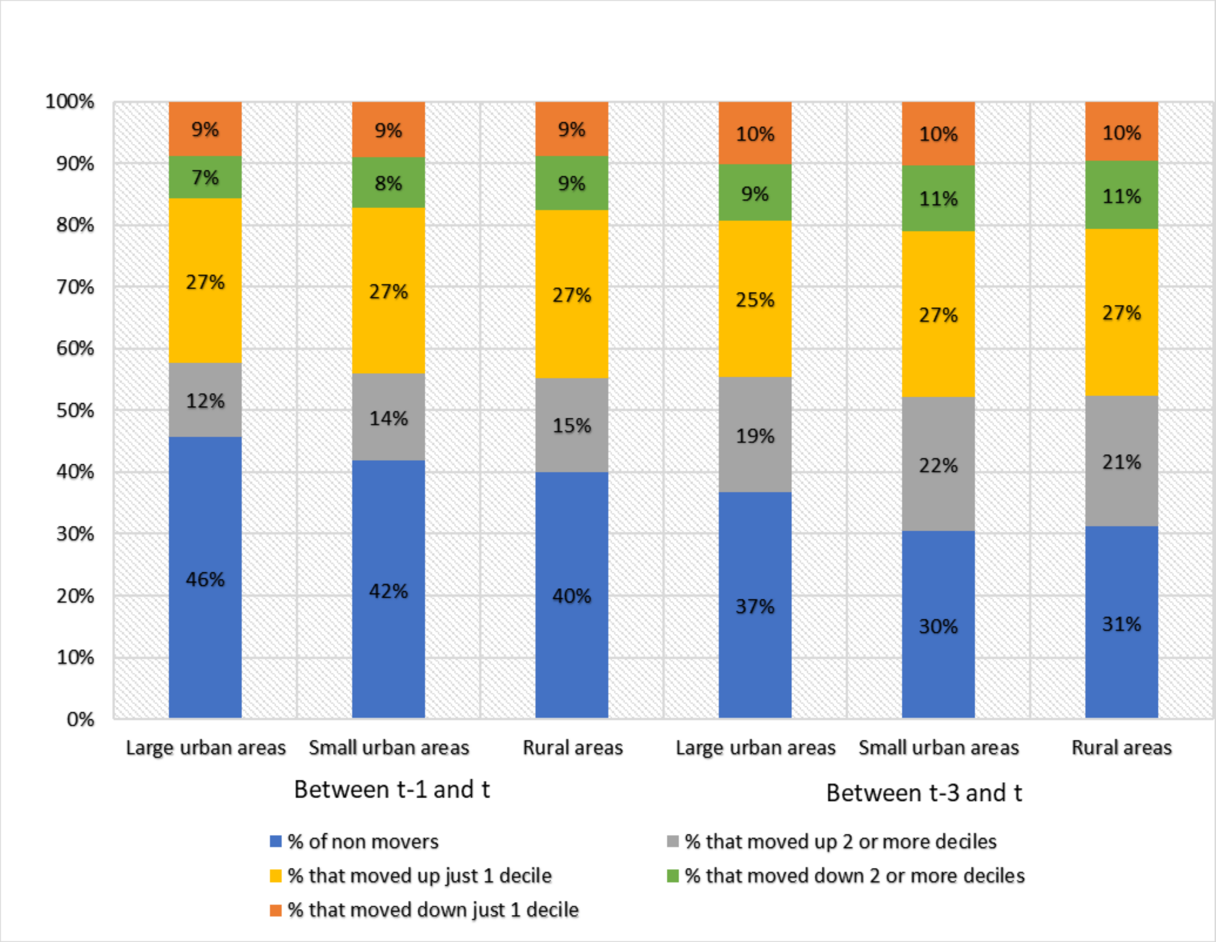


Figure 11: Income mobility in Spain by degree of urbanisation

Figure 11 shows that the level of income mobility for the two periods differs according to the degree of urbanisation of the residential region. Overall, we observe a higher level of both upward and downward income mobility for less densely populated areas. For example, the percentage of workers that moved up two or more deciles in rural areas is 15% (21%) compared to 12% (19%) in large urban areas between t-1 and t (t-3 and t). The same is true for downward mobility. This is an interesting result given the common perception that cities, in particular, large cities offer great opportunities to "climb the socioeconomic ladder".

Co-funded by the Horizon 2020
Framework Programme of the European Union

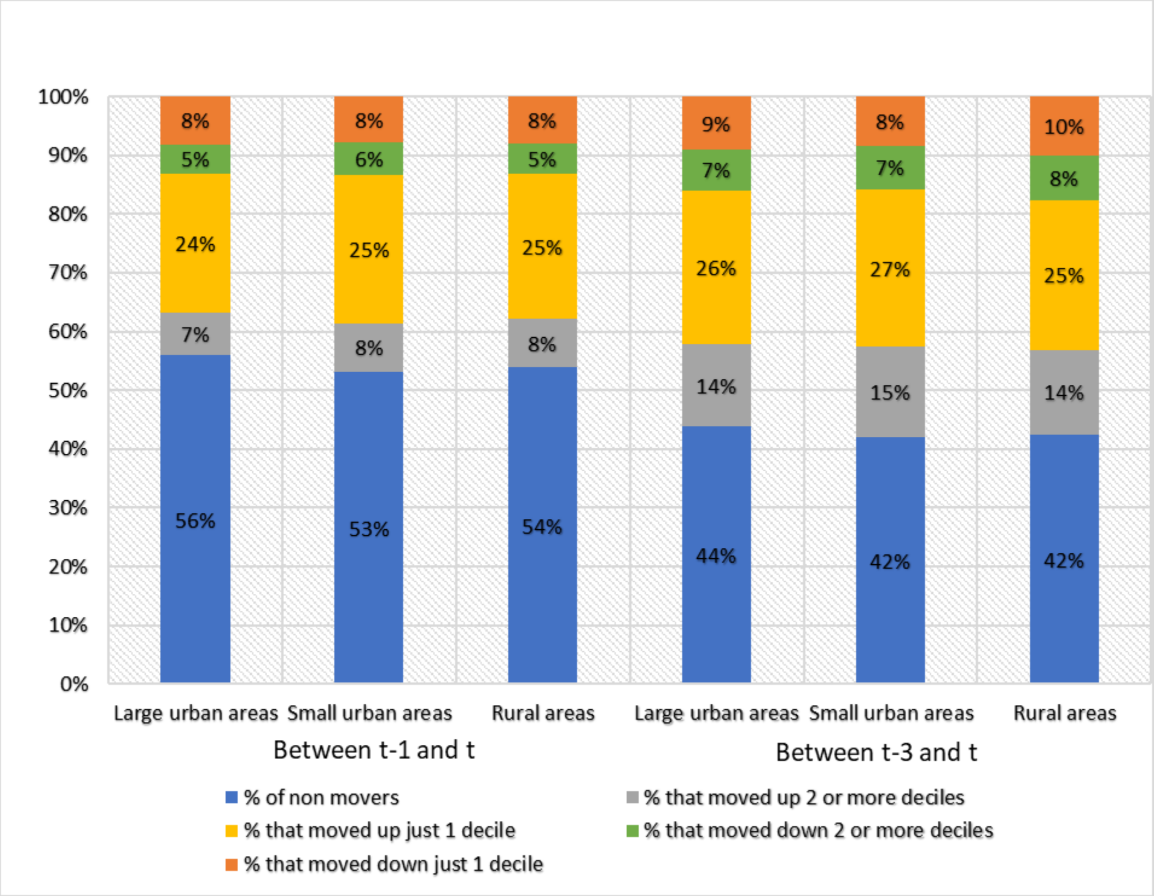


Figure 12: Income mobility in France by degree of urbanisation

Likewise, for Spain, both time intervals varies between different levels of population density in France, and appears to be higher for less densely populated areas. We can also notice that overall, income mobility is higher across all 3 regions in Spain than in France.

Co-funded by the Horizon 2020
Framework Programme of the European Union

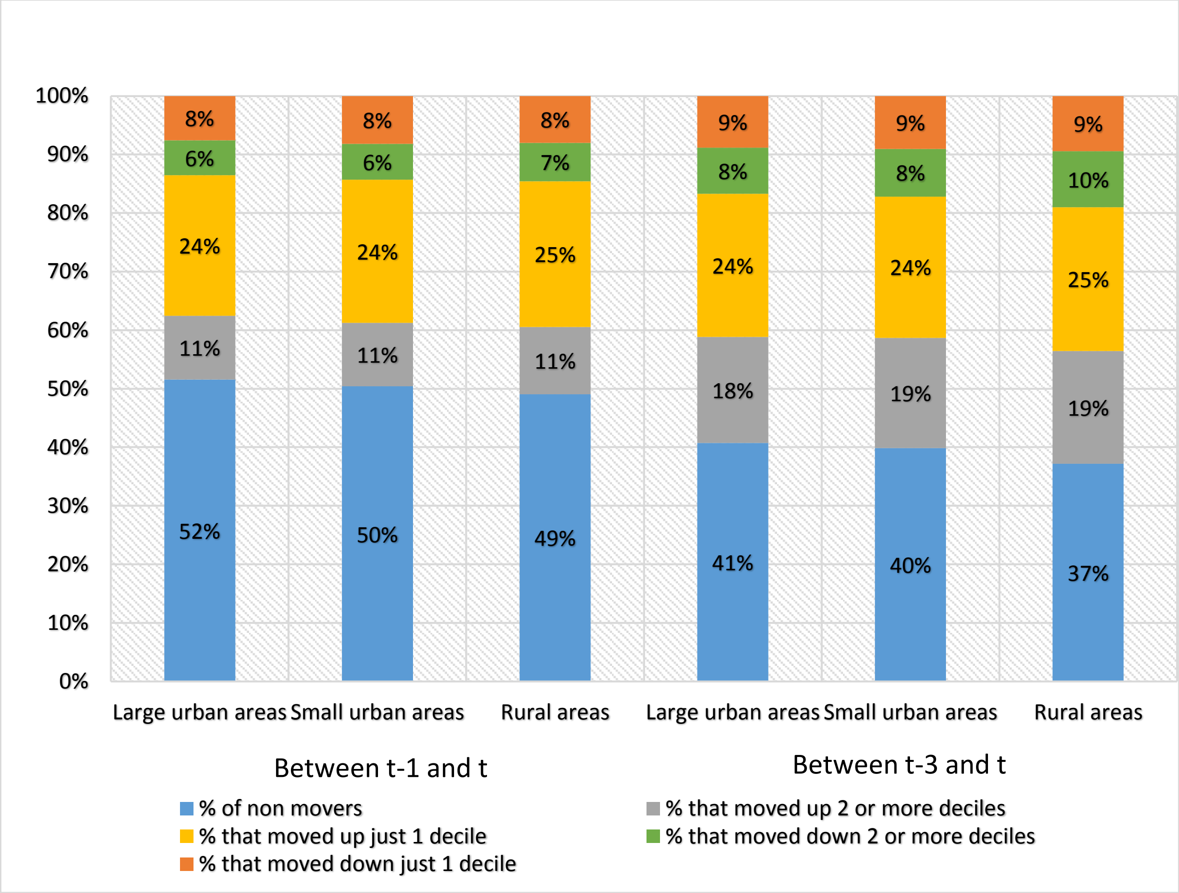


Figure 13: Income mobility in Finland by degree of urbanisation

Likewise, Figure 13 shows that income mobility for both time intervals varies between different levels of population density in Finland and appears to be higher for less densely populated areas.

- **NUTS2 regions analysis**

We now discuss income mobility at the NUTS2 level within each country. Each pair of Figures corresponds to transitions between periods t-1 and t, and between t-3 and t, respectively, for each country.

Co-funded by the Horizon 2020
Framework Programme of the European Union

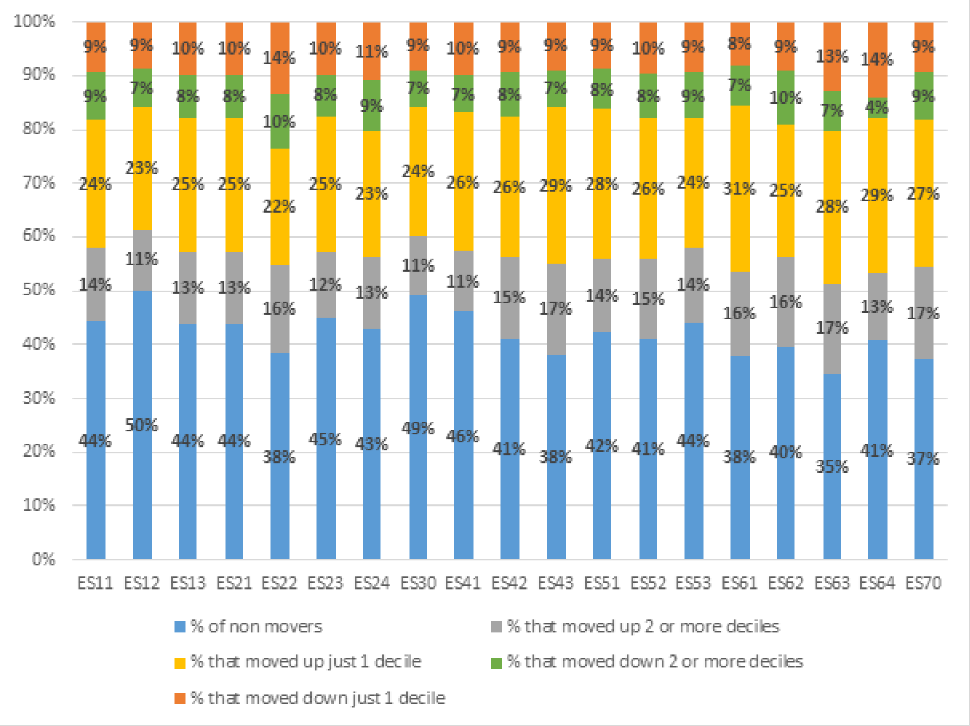


Figure 14: Income mobility in Spain between t-1 and t across NUTS2 regions

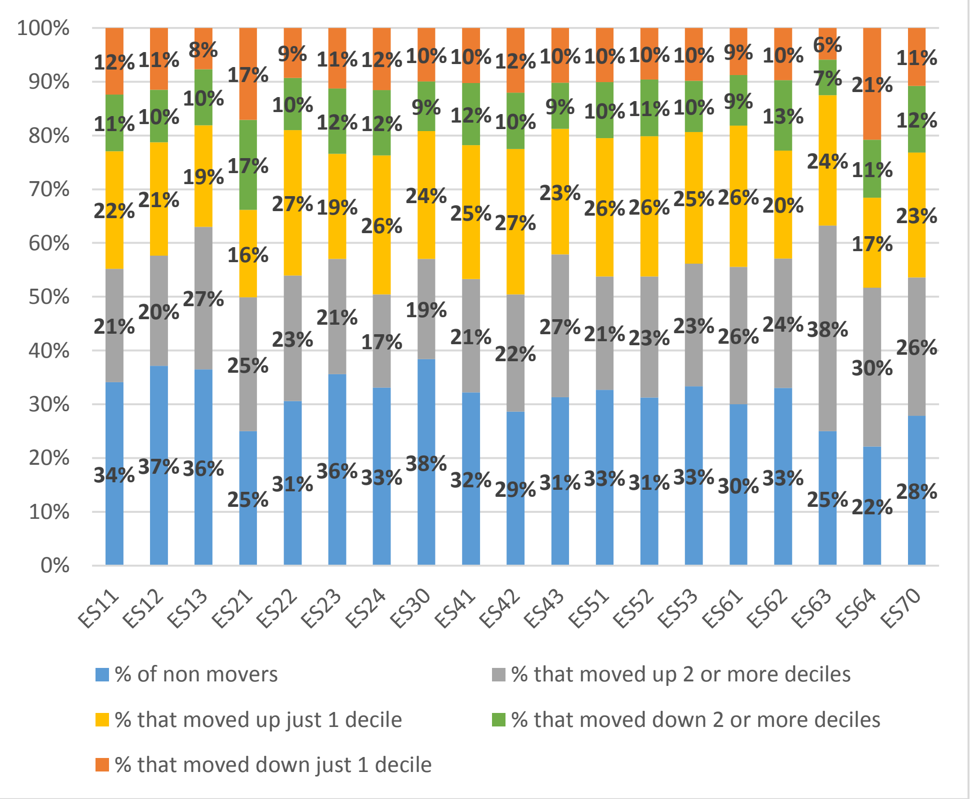


Figure 15: Income mobility in Spain between t-3 and t across NUTS2 regions

Co-funded by the Horizon 2020
Framework Programme of the European Union

There is considerable variation in the extent of income mobility across regions, with the higher percentage of non-movers ranging between 35% and 50% between t and t-1 for Asturias and Ceuta, respectively. Between t-3 and t-1, the tendency of higher overall mobility is preserved, where now the Comunidad de Madrid has the highest percentage of non-movers, and the lowest percentage of individuals that moved both up and down 2 or more deciles. By contrast, the cities of Ceuta and Melilla were the regions with highest mobility both up and down 2 or more deciles in the gross income distribution.

The following scatter plot shows the association between NUTS2 region population and the degree of upward mobility (i.e. the percentage of individuals that moved up one and more deciles the income distribution) between t-1 and t. The Pearson correlation coefficient is very small (-0.022) but statistically significant. The relationship between the two variables does not, however, appear to be linear. A similar relation was found for degree of urbanisation, whereby mobility appears to be slightly higher for less densely populated areas.

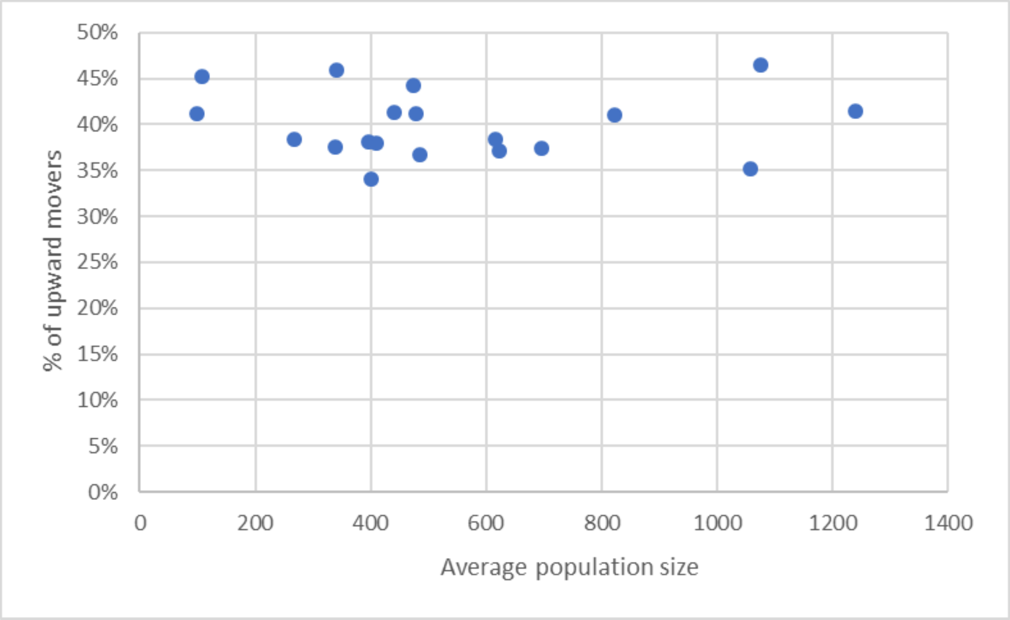


Figure 16: Degree of upward mobility and NUTS2 population size in Spain

Co-funded by the Horizon 2020
Framework Programme of the European Union

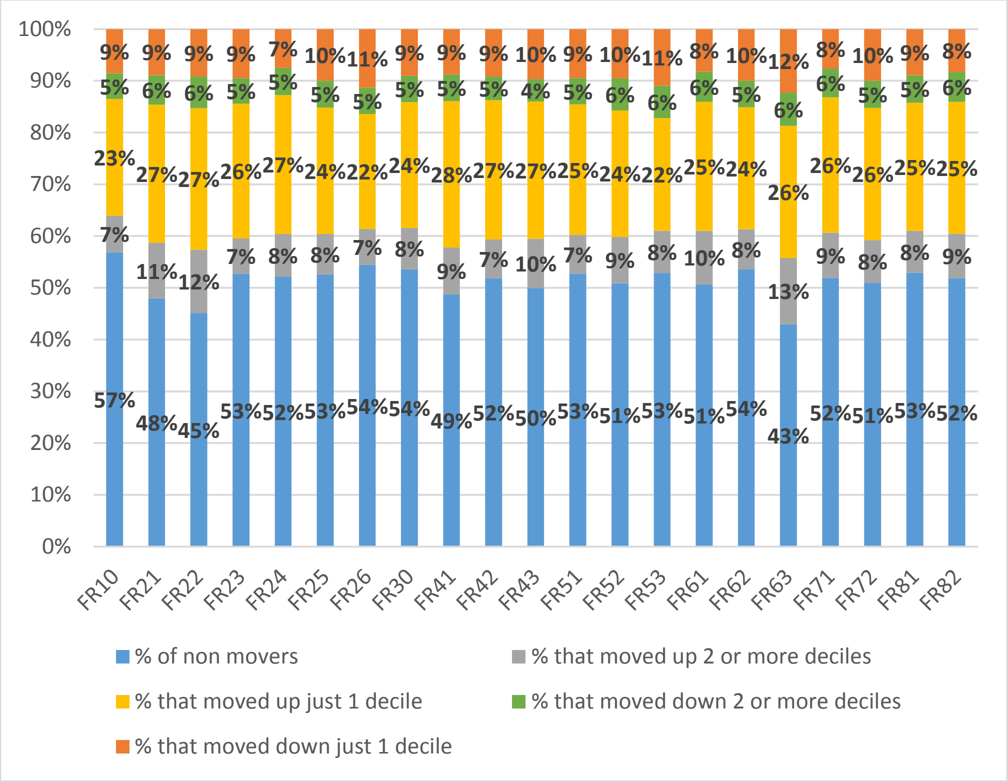


Figure 17: Income mobility in France between t-1 and t across NUTS2 regions

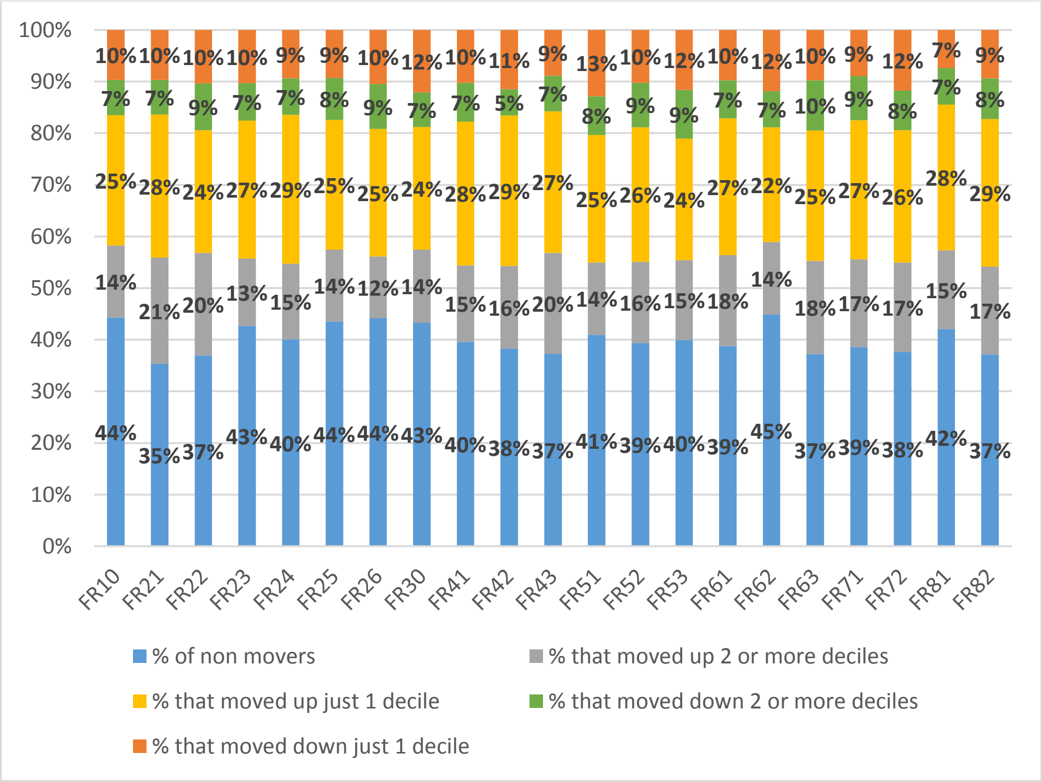


Figure 18: Income mobility in France between t-3 and t across NUTS2 regions

The highest percentage of non-movers was in Île de France between t-1 and t, with a higher proportion upwards in the gross income distribution, whereas the highest mobility between consecutive periods was registered in Limousin, where most of the shift was also upwards. Notice that mobility across most regions between consecutive periods was lower than the average for the whole country. This had to do with the fact that the suburban area of Paris accounts for 17% of the individuals. If we add Nord-Pas-de-Calais and Pays de la Loire, which are amongst the regions with highest percentage of non-movers, we get almost 21% of the individuals across the 21 regions. Therefore, income mobility is heterogeneous across regions, in strong correlation with the degree of urbanisation analyzed previously.

The picture is pretty much the same between t-3 and t in analogy to the comparison across different degrees or urbanisation. Mobility is overall higher but does not change significantly across regions relative to mobility between two consecutive years.

The following scatter plot shows the association between NUTS2 region population and the level of upward mobility between t-1 and t. The Pearson correlation coefficient is -0.36 and statistically significant, which corroborates our previous findings that mobility is relatively higher for less densely populated areas.

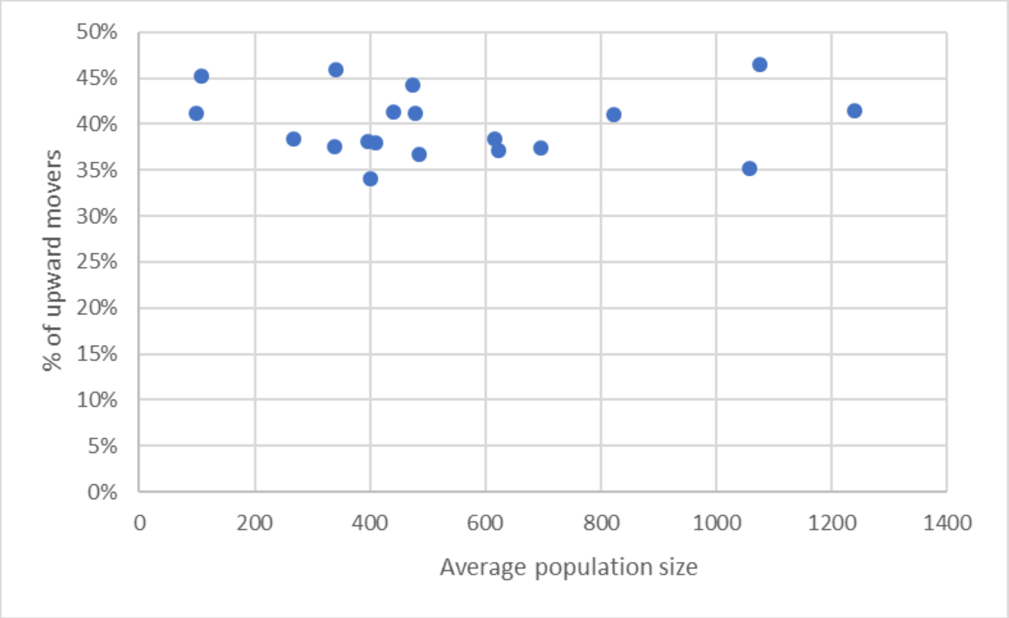


Figure 19: Degree of upward mobility and NUTS2 population size in France

Co-funded by the Horizon 2020
Framework Programme of the European Union

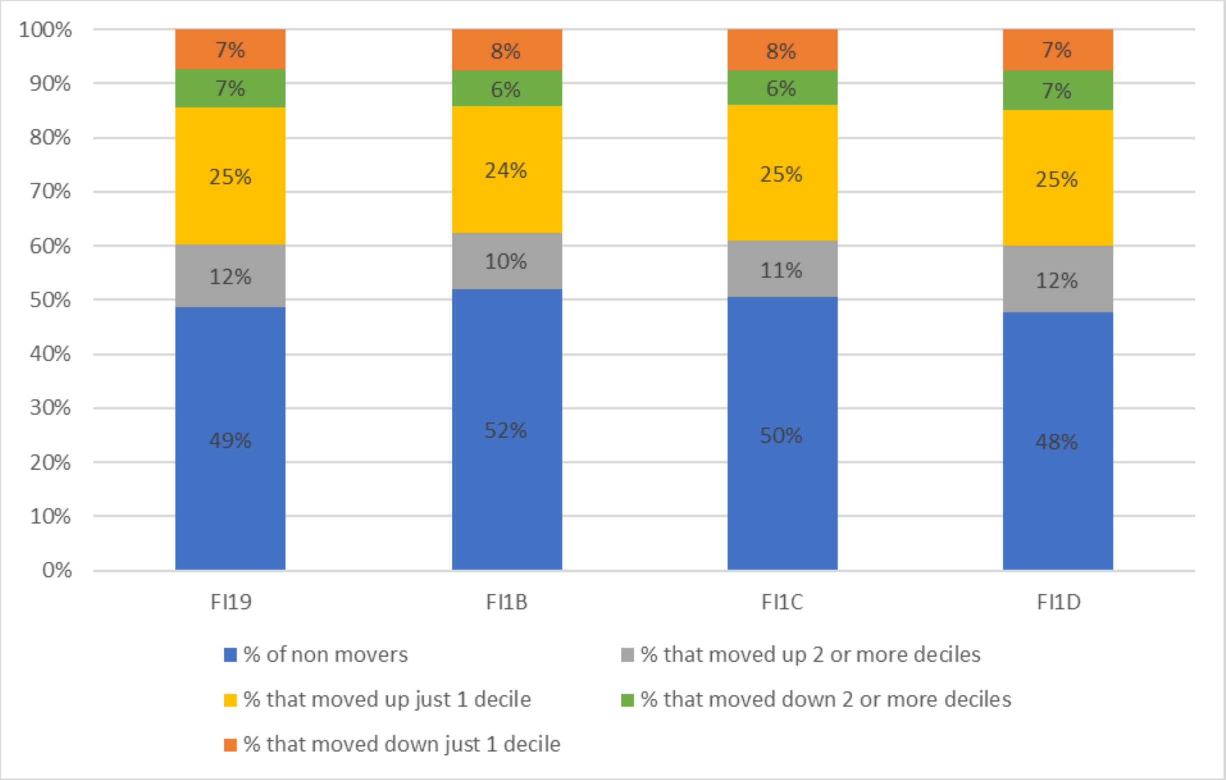


Figure 20: Income mobility in Finland between t-1 and t across NUTS2 regions

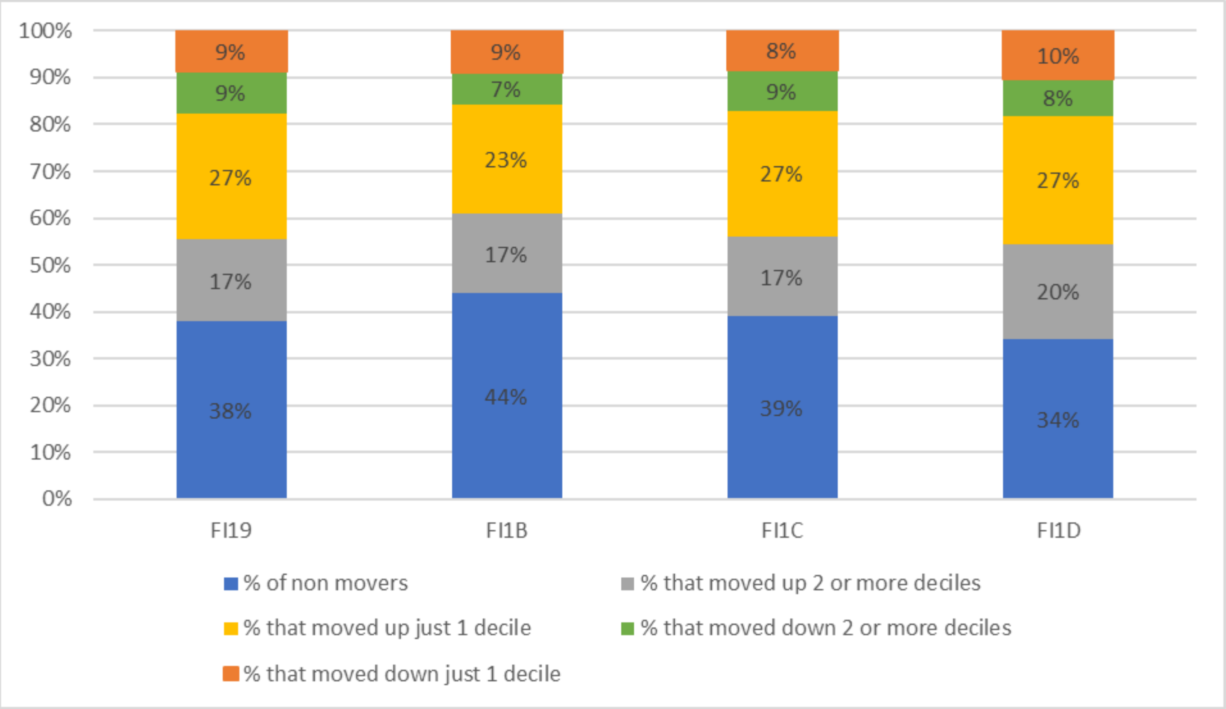


Figure 21: Income mobility in Finland between t-3 and t across NUTS2 regions

Finland reports information at the regional level according to both the old 2006 and the new 2010 NUTS-region codes. For instance, the new region FI1D is the simple aggregation of the old regions FI1A and FI13. Therefore, we can simply merge the latter codes and add to the former. However, the new regions FI1C and FI1B arose from the disaggregation of the old region FI18, which means that we cannot split the old code because we do not have information regarding which shares of the observations in FI18 went to either FI1B or FI2C (López-Cobo, 2016). This implies that we can only safely use years for which no observations were reported for the old region FI18. As a result, the analysis at the regional level is confined to the period 2011-2016.

The highest percentage of non-movers was in Helsinki-Uusimaa both between t-1 and t and between t-3 and t. Comparing the different time intervals, the increase in mobility was higher in Pohjois- ja Itä-Suomi compared to Länsi-Suomi.

When comparing the results across countries, we can say that overall we observe that there appears to be greater variation in the level of mobility across regions in France, followed by Spain, and Finland. However, discrepancies in upward mobility (1 or more deciles) across regions were very high in Spain and France (difference of 12 percentage points between extremes) and lowest in Finland (difference of 4 percentage points between extremes). There is some indication of a negative correlation between regions' population size and the degree of upward mobility in all countries, which corroborates the previous findings regarding mobility by degree of urbanisation.

2. Analysis of income inequality

The analysis focuses on four main measures: Gini Index (GI), income ratio between the 90th and 10th percentiles (P90/P10), income ratio between the 90th and 50th percentiles (P90/P50), and the income ratio between the 50th and 10th percentiles (P50/P10). The Gini Index is the measure of income inequality based on the comparison of cumulative proportions of the population against cumulative proportions of their income. Technically, it is defined as a ratio reporting to the area between the Lorenz curve, which graphs the income share owned by the bottom percentiles of the population, and the uniform distribution (perfect equality) line, ranging from 0 (perfect equality) to 1 (perfect inequality). The income share ratio P90/P10 is the income share ratio between the 10% highest income earners and the 10% lowest income earners. The income share ratio P90/P50 is the income share ratio between the 10% richest and those below the median of the income distribution. Finally, the income share ratio P50/P10 is the income share ratio between the lower median and the 10% poorest.

- **National level analysis**

The results for Spain are reported in Table 26. The Gini index is slightly lower (about one percentage point) than the ones reported for household equivalised disposable income by Eurostat[17], but the trend of increasing inequality is the same. However, one should note that the reference population underlying the analyses is different; the Eurostat report focuses on households whereas we focus on individuals. Moreover, the values from Eurostat pertain to equivalised disposable income whereas we focus on income of employed individuals. Hence, it does not take into account factors such as self-employed income which is prone to higher variability and thus likely bears a negative impact on income inequality. Therefore, one can say that our results still reveal a considerable robustness to those reported by Eurostat. Regarding the income ratios, the 10% richest workers earn 7 to 8 times more than the 10% poorest workers. We can also infer from the table that most inequality in Spain is observed in the lower median compared to the 10% poorest income earners.

Table 26: Summary measures of income inequality in Spain

year	P50/P10	P90/P50	P90/P10	Gini
2005	8.96	0.82	7.40	0.31
2006	8.73	0.80	7.05	0.30
2007	8.80	0.78	6.93	0.30
2008	8.71	0.81	7.08	0.31
2009	8.85	0.84	7.47	0.32
2010	8.99	0.89	8.01	0.33
2011	8.86	0.89	7.97	0.33

[17] http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di12.

Co-funded by the Horizon 2020
Framework Programme of the European Union

year	P50/P10	P90/P50	P90/P10	Gini
2012	9.02	0.88	8.0	0.33
2013	8.93	0.90	8.09	0.33
2014	8.96	0.91	8.17	0.33
2015	8.99	0.93	8.42	0.34
2016	8.72	0.92	8.05	0.34

Considering the results for France in Table 27, we observe that the income ratios have remained constant over the years. The ratio between the 10% richest and the 10% poorest is on average 5.5. The Gini index has also remained almost invariant at 0.27 throughout the years, slightly lower than the ones reported for equivalised disposable household income, which reported just a slightly increase in income inequality regarding the latter measure. The share of the 50% poorest is around 7.5 higher than the share of the 10% poorest.

Table 27: Summary measures of income inequality in France

year	P50/P10	P90/P50	P90/P10	Gini
2006	7.47	0.72	5.37	0.27
2007	7.57	0.70	5.27	0.27
2008	7.54	0.72	5.45	0.27
2009	7.59	0.73	5.52	0.27
2010	7.66	0.72	5.50	0.27
2011	7.73	0.71	5.49	0.27
2012	7.62	0.72	5.45	0.27
2013	7.67	0.70	5.40	0.26
2014	7.62	0.70	5.34	0.26
2015	7.71	0.72	5.52	0.27
2016	7.65	0.71	5.43	0.27

The results for Finland are shown Table 28. The Gini index is slightly lower (about one percentage point) than the ones reported for equivalised disposable income by Eurostat. Regarding the income ratios, the 10% richest earn 4 to 4.5 times more than the 10% poorest, values that are lower than the income share ratio of 5.5 reported for 2013 in the OECD repor (2015) using equivalised disposable income as reference. Moreover, the 10% richest in Finland earn 60% of the 50% poorest income share, presenting the lowest inequality in the upper median of all the four countries.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 28: Summary measures of income inequality in Finland

year	P50/P10	P90/P50	P90/P10	Gini
2005	6.82	0.58	3.96	0.22
2006	6.88	0.60	4.13	0.23
2007	6.92	0.61	4.23	0.23
2008	6.88	0.62	4.25	0.23
2009	7.06	0.63	4.42	0.24
2010	7.01	0.62	4.36	0.24
2011	7.05	0.60	4.25	0.24
2012	7.14	0.61	4.32	0.24
2013	7.15	0.62	4.43	0.24
2014	7.15	0.62	4.41	0.24
2015	7.23	0.62	4.50	0.24
2016	7.21	0.62	4.47	0.24

When comparing the results across countries, we can say that inequality in the lower median of the income distribution compared to inequality between the 10% richest and 10% poorest is much more pronounced in France compared to Spain. The lower inequality is more pronounced when comparing the ratio between highest and lowest income earners, which is highest for Spain, followed by France and Finland. Moreover, the 10% richest in Finland earn 60% of the 50% poorest income share, presenting the lowest inequality in the upper median of all the three countries.

- **Degree of urbanisation analysis**

The following figures show the evolution of the Gini index for the countries under study by type of residential area defined according to degree of urbanisation.

Co-funded by the Horizon 2020
Framework Programme of the European Union

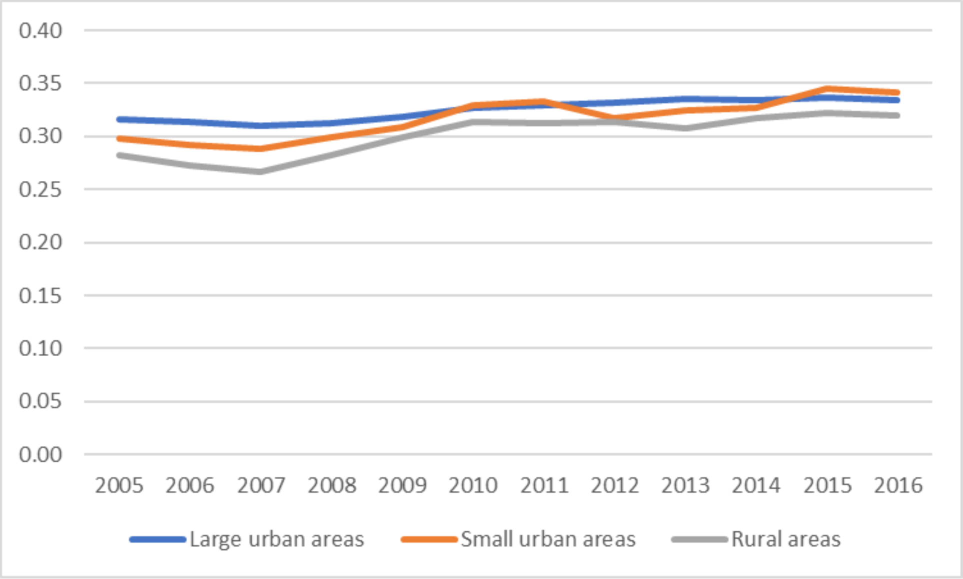


Figure 22: Gini index by degree of urbanisation in Spain

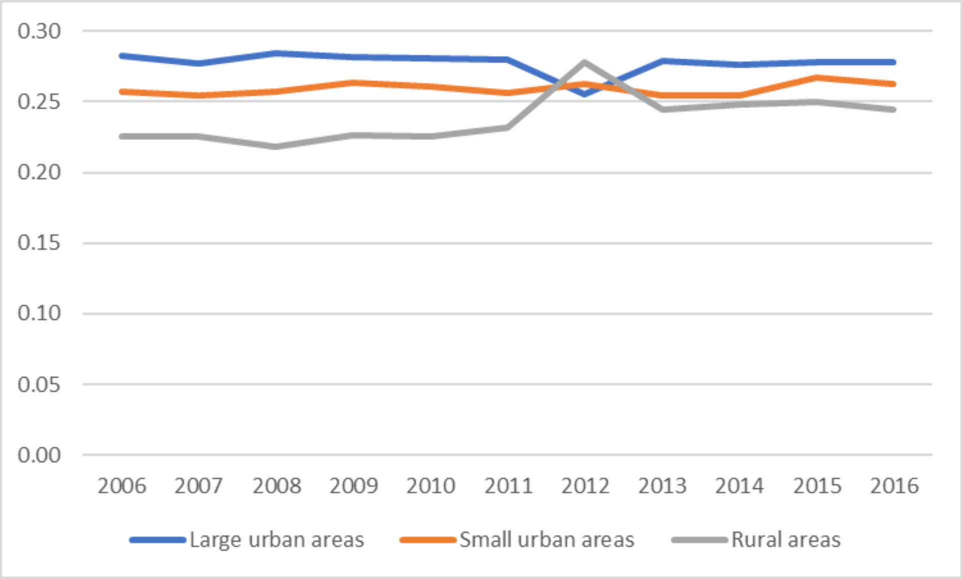


Figure 23: Gini index by degree of urbanisation in France

Co-funded by the Horizon 2020
Framework Programme of the European Union

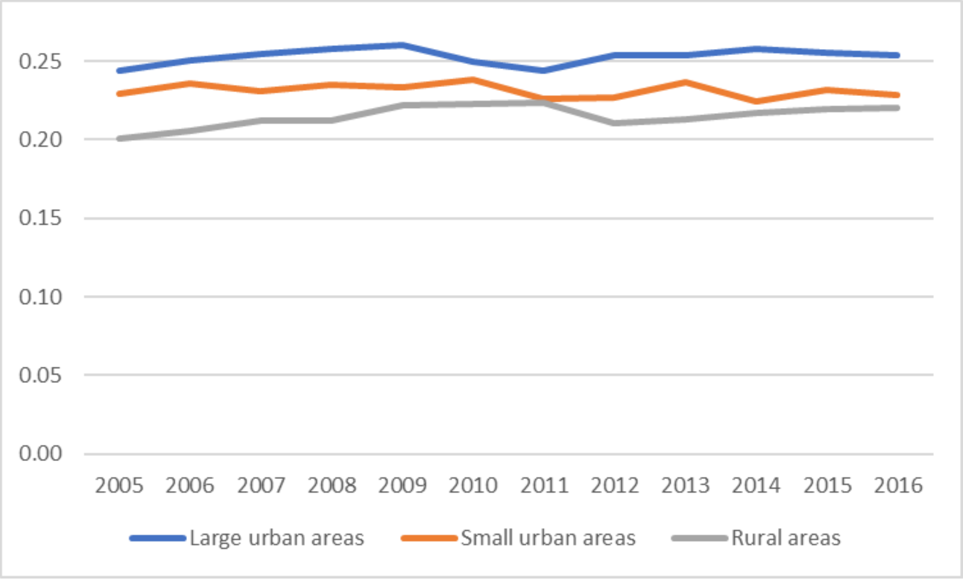


Figure 24: Gini index by degree of urbanisation in Finland

Data for Spain shows that income inequality is higher for large urban areas before 2010, but has become higher in small urban areas compared to large urban years as of 2014 (and also between 2010 and 2011). The trend for the evolution inequality is invariant across degrees of urbanisation, with more pronounced fluctuations in small urban areas. Overall, inequality has risen in accordance with the results at country level and in accordance with the report by Eurostat (2018).

Data for France shows that inequality was overall higher in more densely populated areas throughout most of the entire period. However, the situation reversed completely in 2012, where the ratio was much higher in rural areas compared to large urban areas; it resumed the previous trend there in after. This case seems particularly awkward and leads to suspicion that the code for degree of urbanisation was switched in the reported data for 2012 between large urban areas and rural areas. In sum, inequality remained relatively invariant, in accordance with the results at the National level. This is due to the fact that it decreased in large urban areas, but this decrease was compensated by a slight increase of inequality in small urban areas and a more pronounced increase in rural areas.

Data for Finland shows that that income inequality is higher in more densely populated areas throughout the entire period. Inequality has risen 2 percentage points at the national level, being far more pronounced in rural areas than large urban areas. Interestingly, though oscillations in small urban areas are less smooth compared to other areas, inequality has decreased slightly by 2016.

All countries noticed a decrease in inequality between 2015 and 2016. Inequality is higher in more densely populated areas throughout the period in France and Finland, although the discrepancies are more pronounced in the former. In Spain, between 2010 and 2011 and after 2014 inequality was higher in small urban areas compared to large urban areas.

We next analyze the income share ratios for the countries by degree of urbanisation over the years. The following figures display relative indices for the different income share ratios for small urban areas and rural areas, where the base value is the income share ratio for large urban areas (i.e., P90/P10 = P90/P50 = P50/P10 = 1, for large urban areas).

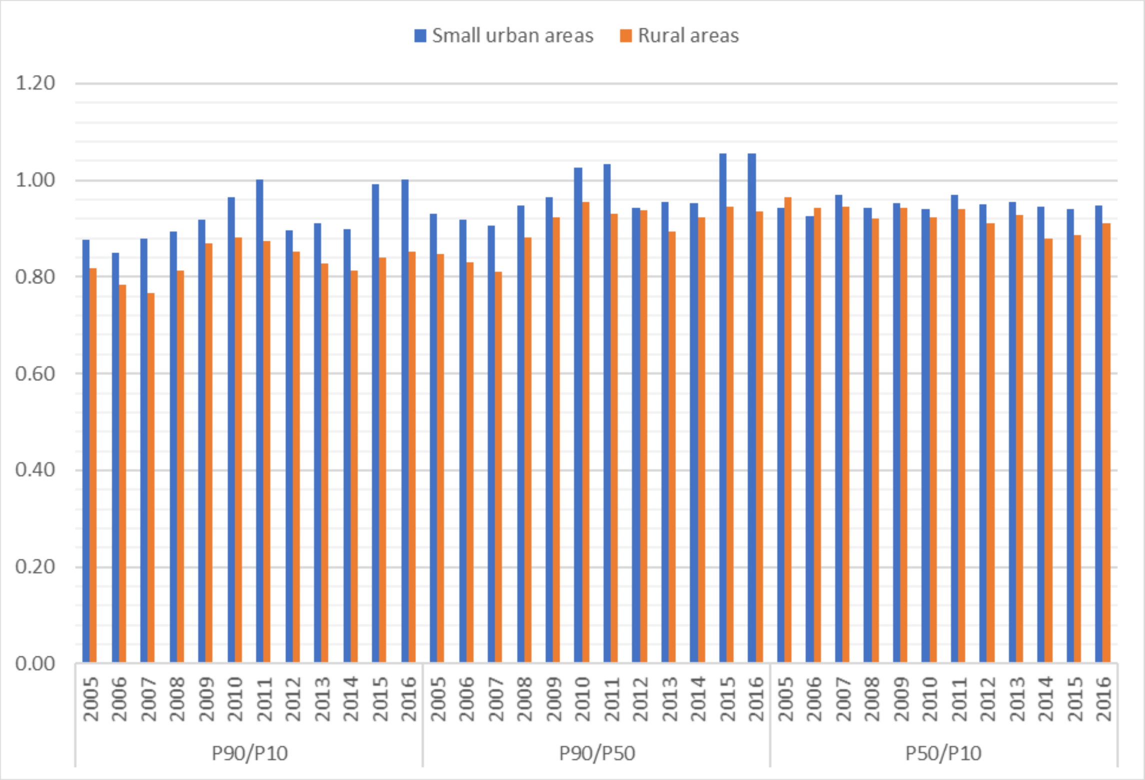


Figure 25: Income ratios for Spain by degree of urbanisation (base=large urban areas)

Co-funded by the Horizon 2020
Framework Programme of the European Union

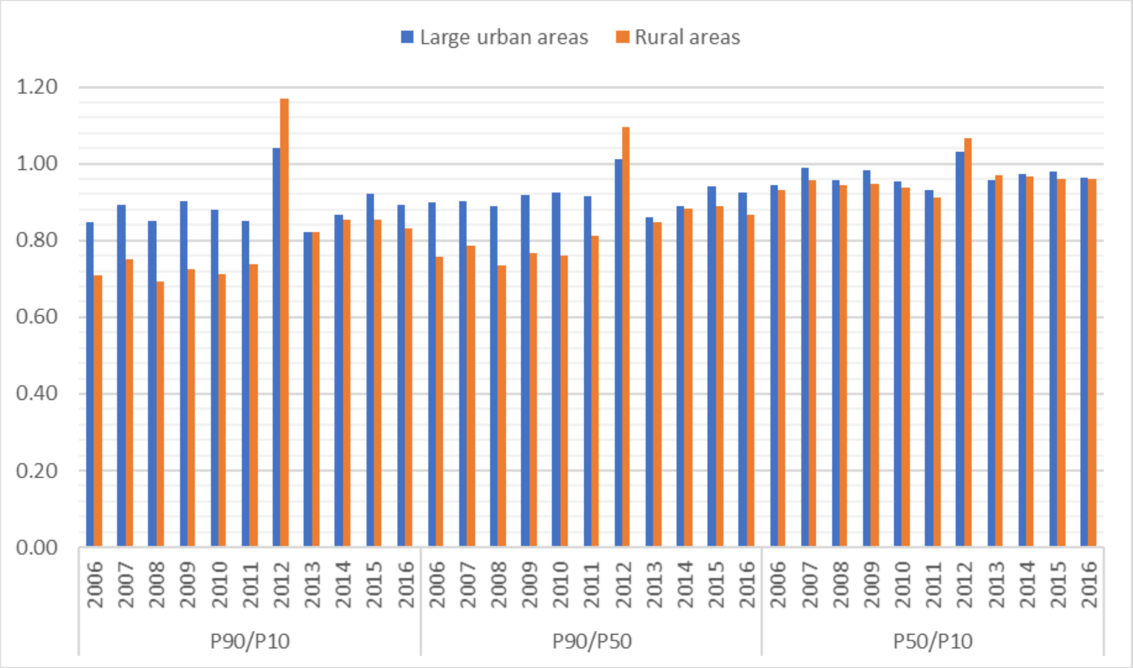


Figure 26: Income ratios in France by degree of urbanisation (base=large urban areas)

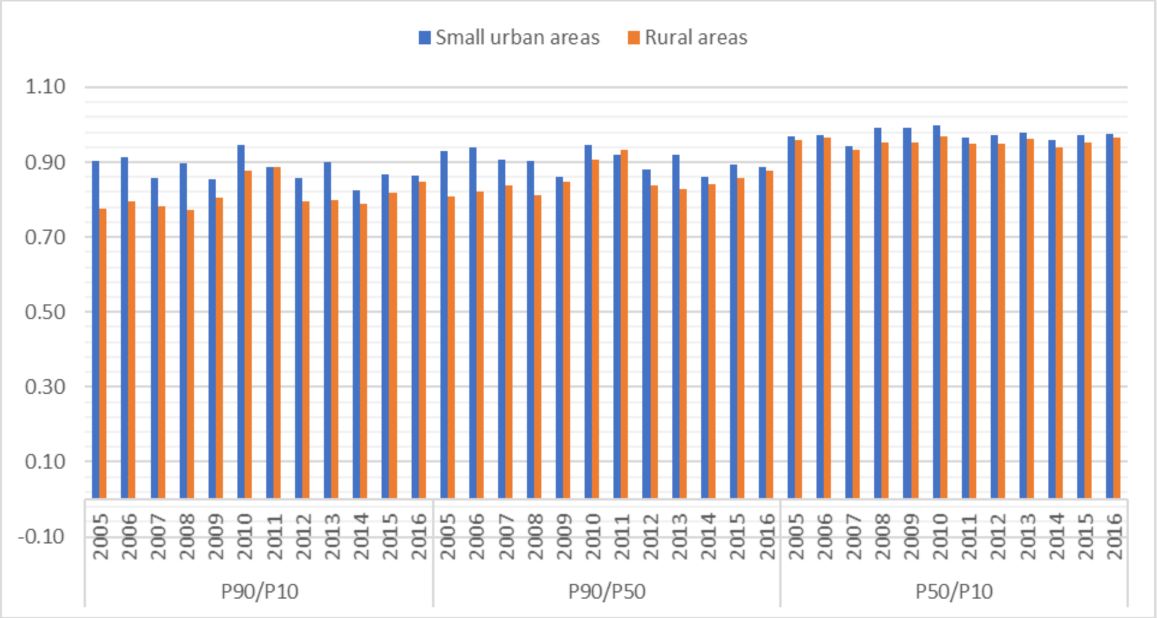


Figure 27: Income ratios in Finland by degree of urbanisation (base=large urban areas)

Data for Spain shows that income inequality is higher in more densely populated areas for most of the period, but this trend is inverted in 2016 regarding the comparison between large and small urban areas. This is especially noticeable when comparing the 10% richest with the 10% poorest across years and degree of urbanisation. However, the 50% poorest earn considerably more than the 10% poorest in large urban areas compared to small urban areas

Co-funded by the Horizon 2020
Framework Programme of the European Union

and rural areas, a trend which hold throughout the entire period, although with noticeable oscillations in magnitude.

Data for France shows that the income shares ratios differ significantly across degree urbanisation until 2011. In 2012 income inequality was completely reversed being highest in rural areas and smallest in large urban areas. After that, the initial trend resumes although in 2013 and 2014 the income shares ratio between the 10% richest and 10% poorest was practically the same in small urban areas and rural areas. The preceding analysis also holds for income shares ratio between the 10% richest and 50% poorest, although the differences in magnitude are much lower.

Data for Finland shows that income inequality is higher for large urban areas, but this discrepancy is more noticeable considering the income share ratios between the 10% poorest and the 10% poorest. The 50% poorest earn considerably more than the 10% poorest in large urban areas compared to small urban areas and rural areas (the latter two being fairly similar regarding this indicator), a trend which hold throughout the entire period, although with noticeable oscillations in magnitude.

Comparing all countries, it is noticeable that the higher discrepancies in inequality across different degrees of urbanisation are observed in the income share ratio between the 10% highest income earners and the 10% lowest income earners. This holds particularly for the case of France.

- **NUTS2 regions analysis**

We now replicate the preceding analysis to the NUTS2 region level. We first discuss the results separately for each country and then provide a cross-country comparison paragraph highlighting the main differences across countries. In the following figures, we report the income share ratios and Gini index as relative indices using as base the values for the capital region for each country (i.e., P90/P10 = P90/P50 = P50/P10 = GI = 1, for the capital region), namely Comunidad de Madrid in Spain, Île de France in France and Helsinki-Uusimaa in Finland.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Figure 28 and Figure 29 refer to Spain. For the sake of exposition, we illustrate our results only for 2 years: 2006 (pre-crisis year), and 2015.18 In 2006, the highest P90/P10 was verified for the Cantabria region, followed by Comunidad de Madrid and Basque Country, whereas the lowest was in La Rioja and Melilla, followed by Comunidad Valenciana and Castilla y Léon. In 2015, the highest P90/P10 was reported in Basque country, although it increased also in Comunidad de Navarra and in the city of Melilla. As for the Gini index, it remained relatively invariant across most regions in both years (although it as increased on average throughout the years), but heterogeneity exists for Cataluña, Illes Balears, Andalúcia, Múrcia and Ceuta, where it was higher compared to other regionsand also where it increased the most.The Gini coefficient also varied a lot in Comunidad Valenciana, La Rioja, and Cantabria

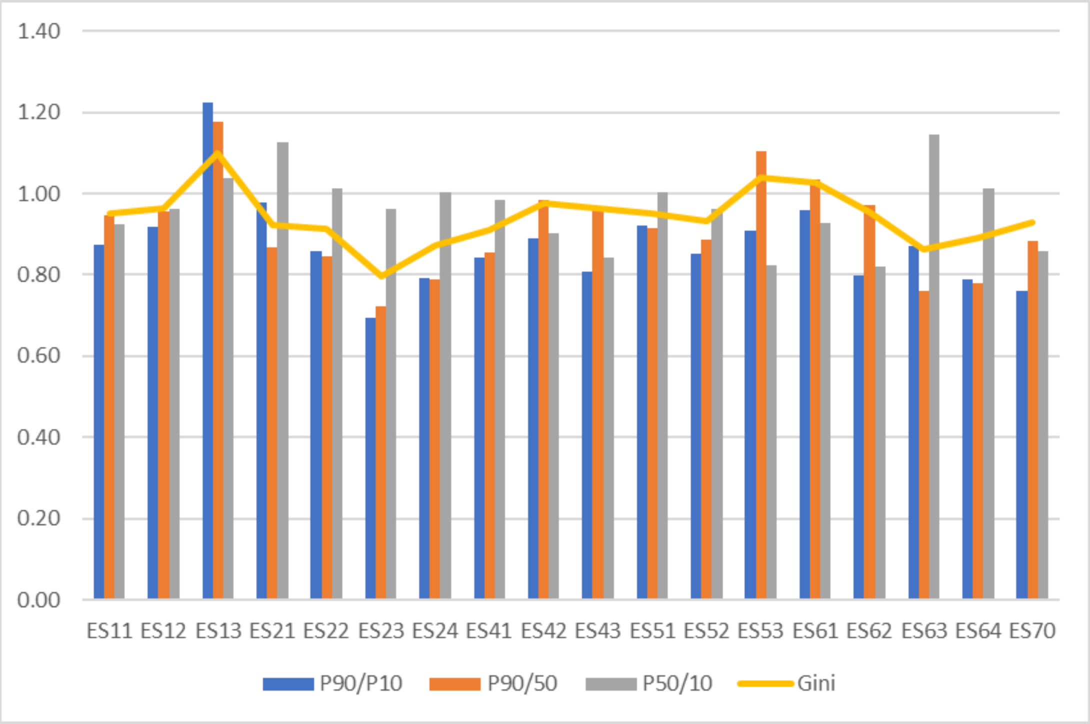


Figure 28: Income ratios and Gini index by region in relation to the capital region, 2006

[18] We avoid 2016 due to possible discrepancies in the number of observations.

Co-funded by the Horizon 2020
Framework Programme of the European Union

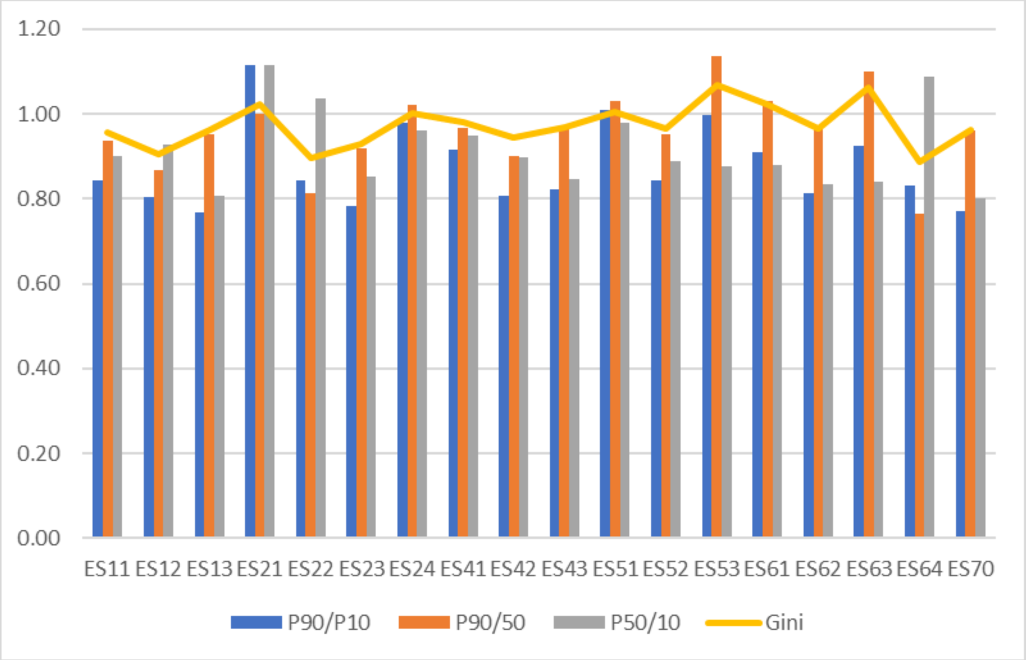


Figure 29: Income ratios and Gini index by region in relation to the capital region, 2015

Figure 30 shows the degree of association between NUTS2 regions' population and the corresponding Gini index over the years for Spain. The correlation coefficient is 0.21 and is statistically significant, suggesting there is a positive association between urbanisation and inequality.

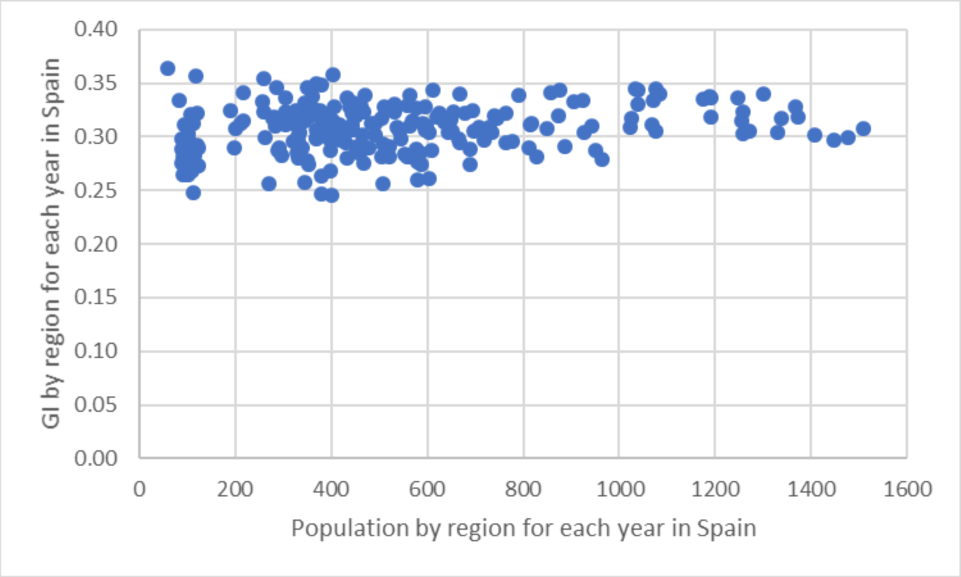


Figure 30: Gini index and NUTS2 region size for each year in Spain

Co-funded by the Horizon 2020
Framework Programme of the European Union

Figure 31 and Figure 32 refer to France. We illustrate our results for 2007 (pre-crisis year), and 2015.[19] In 2007, the highest income share ratio between the 10% richest and the 10% poorest was reported in Île de France, followed by Midi-Pyrénées and Rhône-Alpes whereas the lowest was in Franche-Comté, Centre (FR), Auvergne and Limousin. In 2015, the highest P90/P10 was again in Île de France, but now followed by Centre (FR), in sharp contrast with 2007. Rhône-Alpes follows, but also Languedoc-Roussillon and Provence-Alpes-Côte d'Azur, which were also relatively lower in 2007, particularly for the former. Overall, we conclude that income inequality in France is just as heterogeneous across different regions as across different degrees of urbanisation. The same can be interpreted from the trend of the Gini coefficient. On the other hand, there seems to be an overall decline in the gap between the income share ratios P90/910 and P50/P10, apparently driven by an increase in the income share of the 10% poorest.

[19] We avoid 2006 and 2016 due to possible discrepancies in the number of observations. There are no observations for the case of France in 2005.

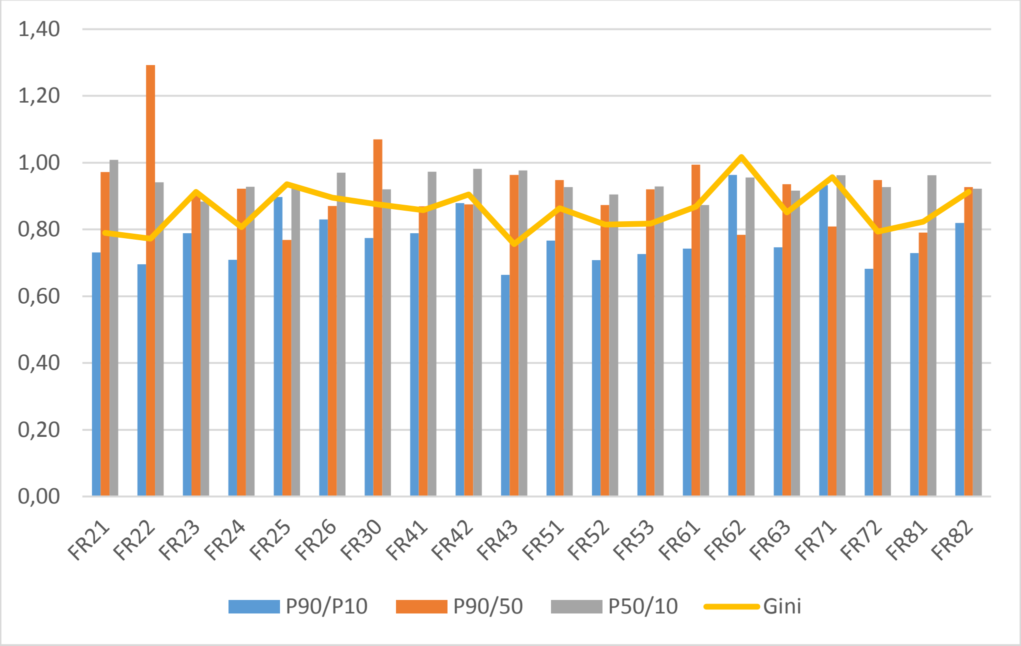


Figure 31: Income ratios and Gini index by region in relation to capital region, 2007

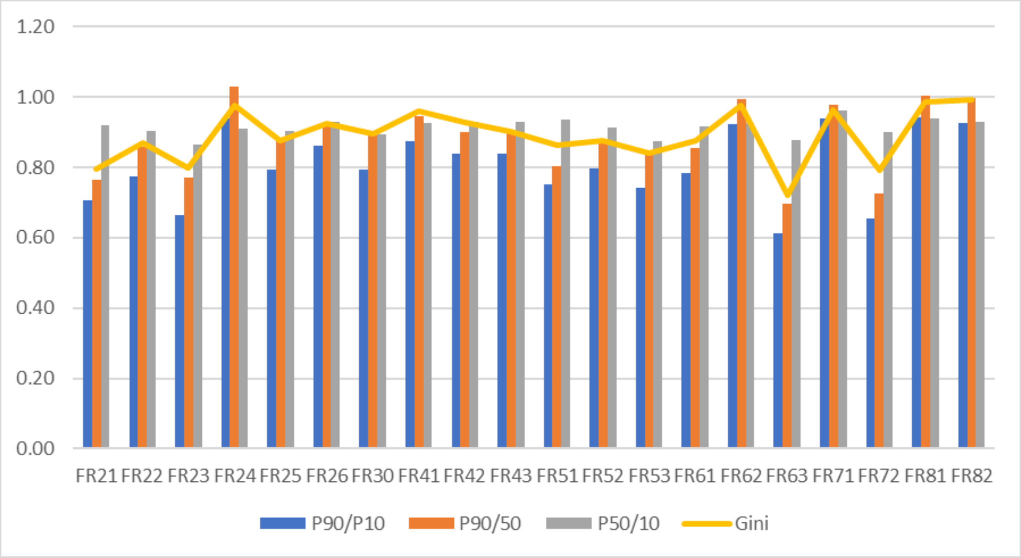


Figure 32: Income ratios and Gini index by region in relation to capital region, 2015

Figure 33 shows that there is a positive association between NUTS2 regions' population and the Gini index for France. The correlation coefficient is 0.53 and is statistically significant, suggesting there is a positive association between urbanisation and inequality. The strength of this association is considerably stronger in France than in Spain.

Co-funded by the Horizon 2020
Framework Programme of the European Union

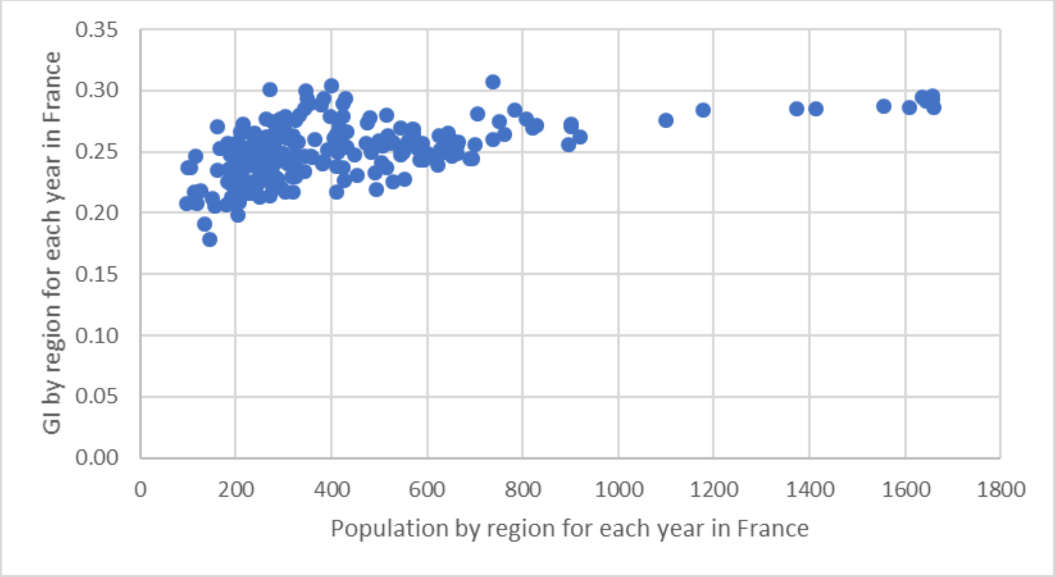


Figure 33: Gini index and NUTS2 region size for each year in France

Figure 34 shows the evolution of the income ratios and Gini index for Finland. The highest income share ratios are observed for the region of Helsinki-Uusimaa, but the difference compared to the other regions for P50/P10 was relatively smaller than the difference for P90/P50, which in turn was smaller than the difference for P90/P10 (with the exception of Länsi-Suomi in 2011). Across the other regions, heterogeneity is small regarding all inequality indicators. Throughout the period, most inequality is observed between the lower median and the 10% poorest. As for the Gini index, shows the prevalence of higher inequality in Helsinki-Uusimaa compared to the rest of Finland. There have been oscillations over time and across regions, though not very pronounced. In fact, whereas inequality was close to that of Helsinki-Uusimaa in Länsi-Suomi in 2011, it decreased until 2016 increased in the other two regions showing evidence of convergence in 2016 between the three regions measured by the Gini coefficient.

Co-funded by the Horizon 2020
Framework Programme of the European Union

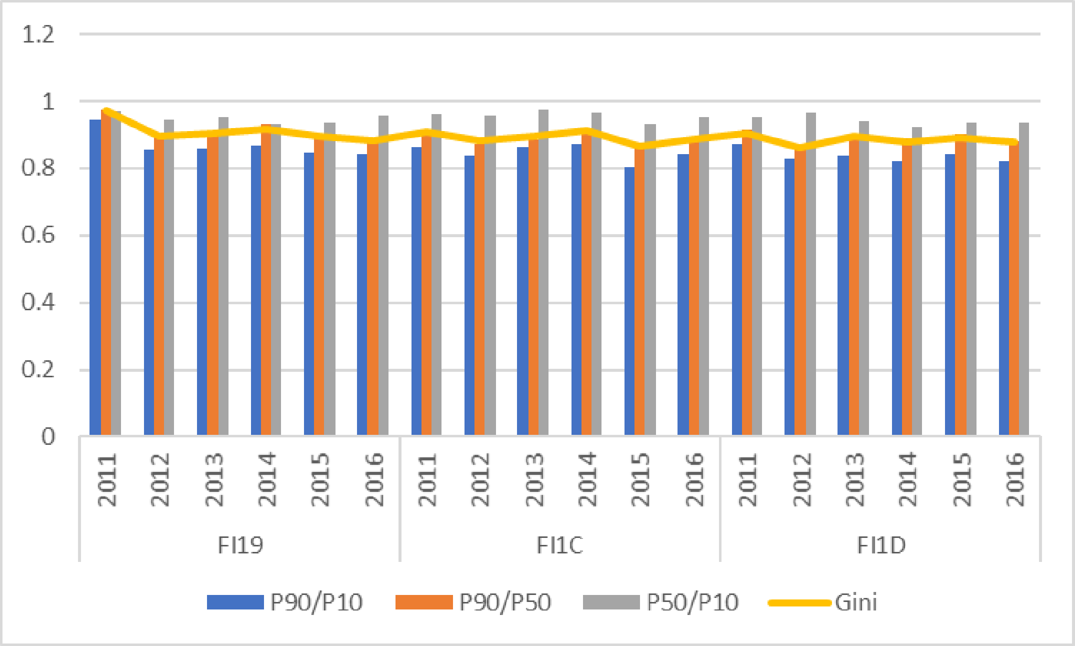


Figure 34: Income share ratios and Gini index in Finland by region (base=capital region)

A comparison between gross and net income inequality

We now compare gross and net income inequality at the national level, by degree of urbanisation, and by NUTS2 region. We exclude Finland due to absence of information regarding net employee income.

- **National level analysis**

Figure 35 illustrates the ratio between the Gini index using net income and the Gini index using gross income for Spain and France. A ratio lower than 1 means that the State is progressive because inequality based on net income is lower compared to gross income. Since gross income here corresponds to net income plus any deductions made by the employee at source, it partially explains the redistributive of taxation. We also emphasize that our focus lies on employee income and not equivalised income for all individuals. Spain exhibits a lower income inequality after deductions, although this progressiveness is more pronounced in Spain. We also observe a steady increase in the redistributive role of worker contributions in Spain. Interestingly though is the case of France, whereby from the Figure it seems that contributions have no redistributive role in France. However, one must be aware that France did not have a

Co-funded by the Horizon 2020
Framework Programme of the European Union

deduction-at-source system by 2016.[20] Moreover, gross labour is net of tax on social contributions, as mentioned in Section 4.1. Gross labour income differs from net labour income in that an imputation method for gross to net conversion is applied to the latter.[21] This helps explain the similarities in the income distributions of both gross and net incomes in France.

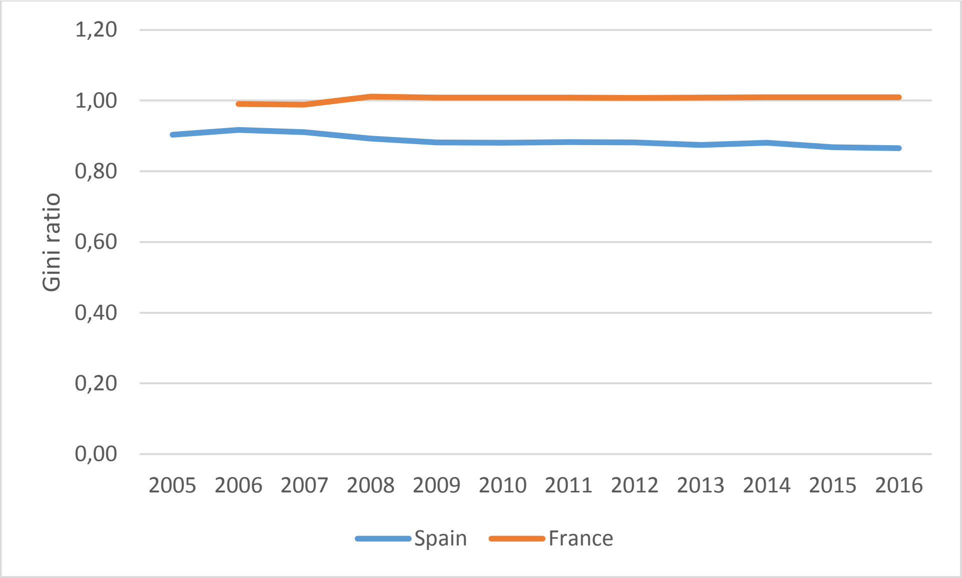


Figure 35: Ratio between the Gini index using net income and gross income

- **Degree of urbanisation analysis**

Figure 36 compares the ratio between the Gini index using net income and the Gini index using gross income across residential areas by degree of urbanisation for Spain, and France. Overall, we conclude that net income inequality in France is just as heterogeneous across the different

[20] This does not forcibly imply that the ratio has to match exactly unity because some exceptions or voluntary contributions could be made in principle.

[21] The corresponding flag value take the value 33, which means that the imputation factor (in percentage) corresponds to the collected value divided by the recorded value. The recorded value is obtained through a statistical model in which parameters are estimated from the data.

Co-funded by the Horizon 2020
Framework Programme of the European Union

regions as gross income inequality and relatively invariant after 2008, with all the ratios very close to 1 just as for the national level.

For the case of Spain, we observe that net income inequality is also lower compared to the case of gross income, for all degrees of urbanisation and following a similar trend over the years. The trend for net income inequality across different degrees of urbanisation measured by the Gini coefficient is qualitatively like the one for gross income. However, the overall decrease in inequality is higher for large urban areas compared to small urban areas, which in turn is higher compared to rural areas.

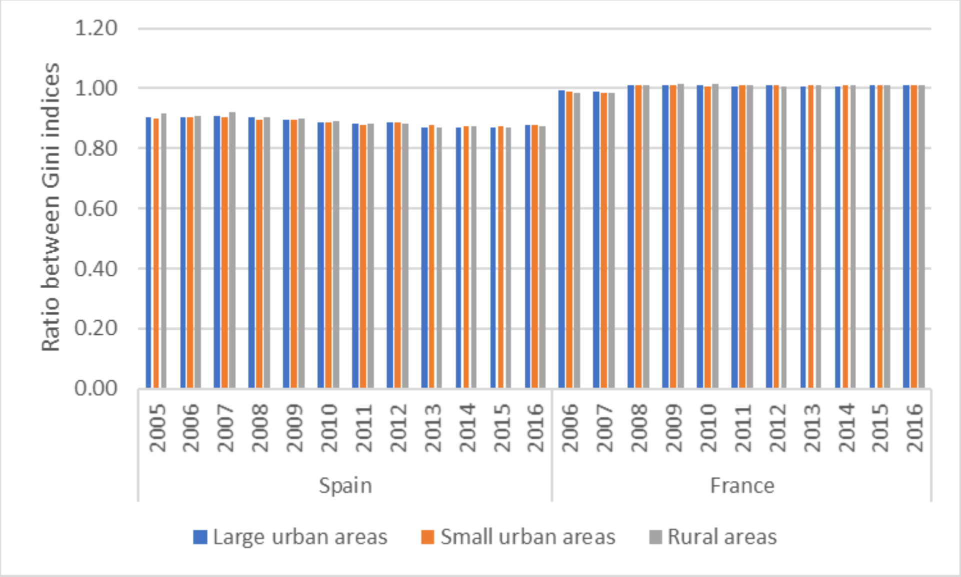


Figure 36: Ratio of Gini indices for net and gross income, by degree of urbanisation

- **NUTS2 regions analysis**

Given the many regions in Spain and France, we present the same pre- and post-crisis two-year comparisons: 2006 and 2015; and 2005 and 2015, respectively.

Figure 37 refers to Spain and shows the difference in magnitude of the Gini index ratio between net and gross values between 2005 and 2015 is more pronounced in Aragón, Comunidad de Madrid, Cataluña, Balearic Islands and Canarias, where progressiveness between those years has improved considerably. However, only in Basque Country and Melilla has the ratio increased, reflecting a decrease in redistribution in those two regions.

Co-funded by the Horizon 2020
Framework Programme of the European Union

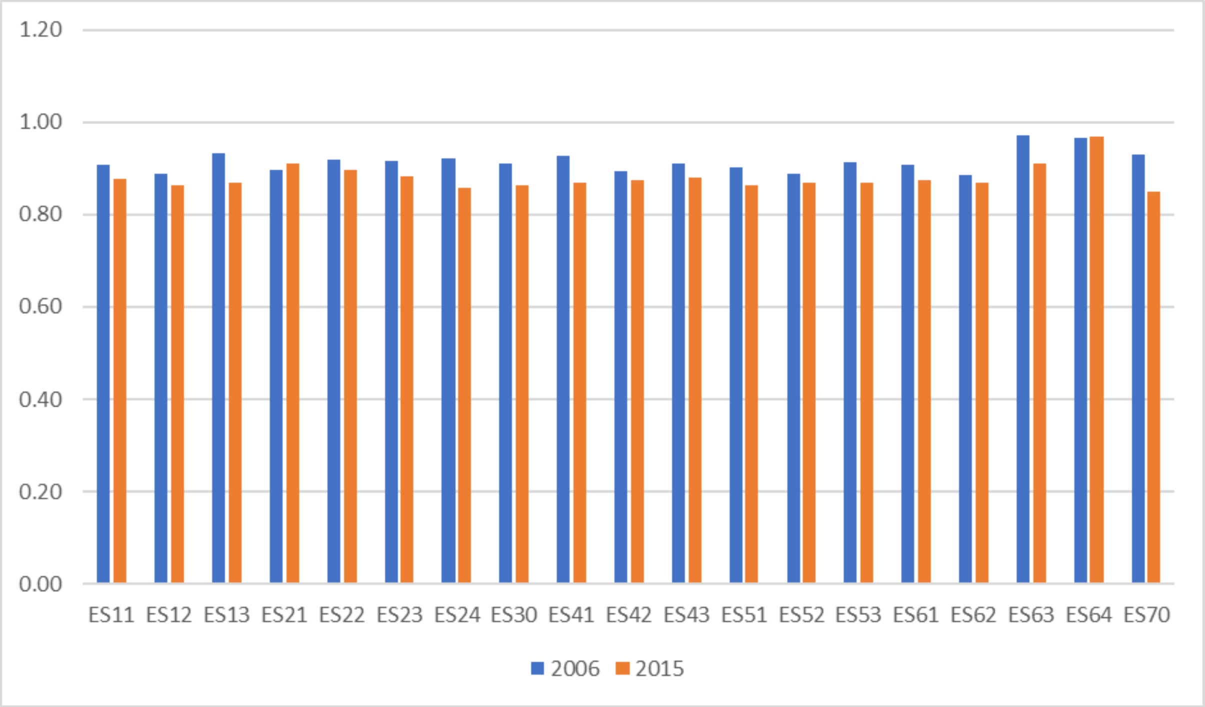


Figure 37: Ratio of Gini indices for net and gross income for Spain, by region

Figure 38 refers to France. As discussed above, there seems to be little (for 2007) to no progressiveness (for 2015) of tax and social contributions for the case of France. This result is somewhat trivial as it could be easily deduced from the Figures at national level and by degree of urbanisation. In 2015 net income inequality was slightly higher than gross income inequality for all regions. However, it can be shown through the data that net income was lower (just

Co-funded by the Horizon 2020
Framework Programme of the European Union

slightly) than gross income for all individuals with the exception of only 163 for which it was the same.

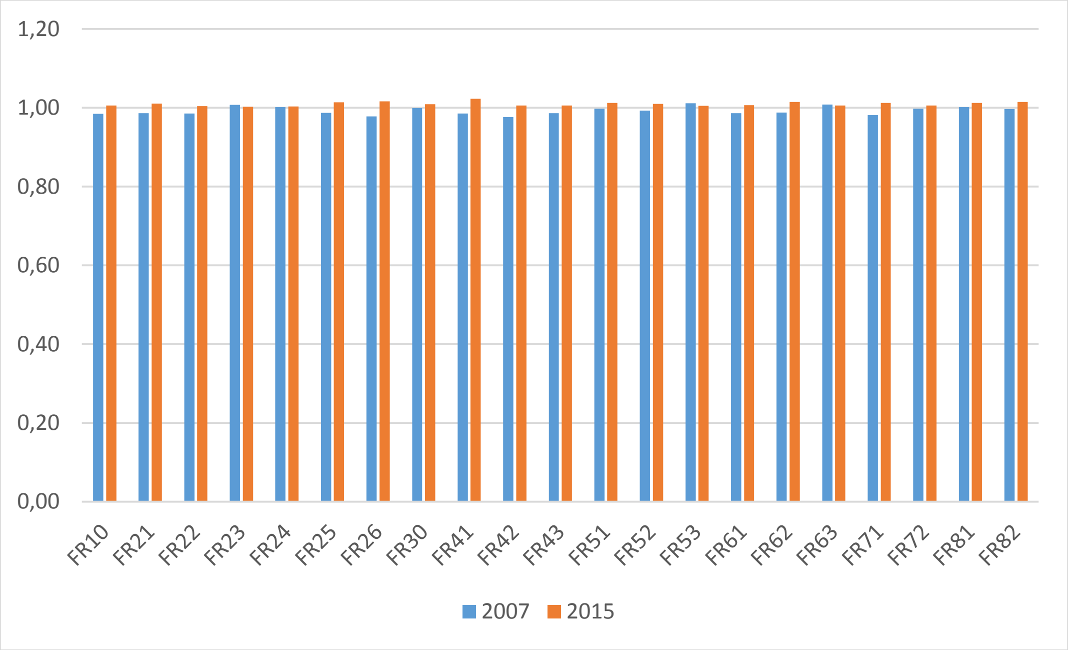


Figure 38: Ratio between Gini indices based on net and gross income for France, by region

Comparing mobility and inequality

While differences between gross and net income do not lead to significant differences in income mobility for most cases, they should have a significant impact on the level of inequality. Similarly, income mobility can be correlated with income inequality. According to Shorrocks (1978) it is expected that higher income mobility implies a lower income inequality for the same reference period.

Figure 39 shows two diagrams pertaining to the Spanish case. To the left, a scatter plot between degree of upward mobility and the Gini coefficient averaged over the reference period. To the right, a scatter plot between the percentage of movers (up and down) the income distribution and the same Gini index. To the left, there is a positive correlation between upward mobility and Gini index, which means that upward mobility seems negatively correlated with lower income inequality. However, this correlation is small (0.044) and is not statistically significant. To the right, by contrast, there seems to be a positive correlation

Co-funded by the Horizon 2020
Framework Programme of the European Union

between higher income mobility and lower income inequality. However, again, the correlation coefficient (-0.22) is also not statistically significant, so no apparent association between mobility and inequality exists for the case of Spain.

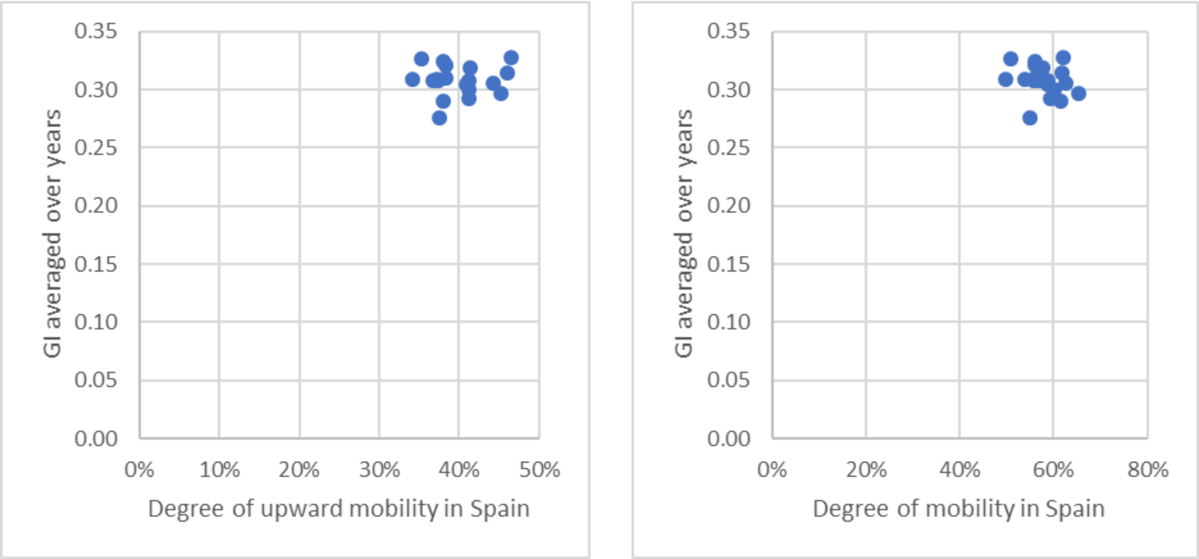


Figure 39: Upward and overall mobility and Gini index for Spain

Figure 40 depicts the same scenario for France. The left diagram shows a positive correlation between higher upward income mobility and lower income inequality. The correlation coefficient is -0.35, indicating that greater inequality is associated with lower income mobility, but is only significant at the 10% level. Likewise, the diagram on the right shows there is a negative association between income inequality and overall income mobility; the correlation coefficient is -0.60 and statistically significant, indicating that greater inequality is associated with overall lower income mobility. The evidence for France also gives support to Shorrocks' hypothesis (Shorrocks, 1978) that higher mobility is associated with lower income inequality.

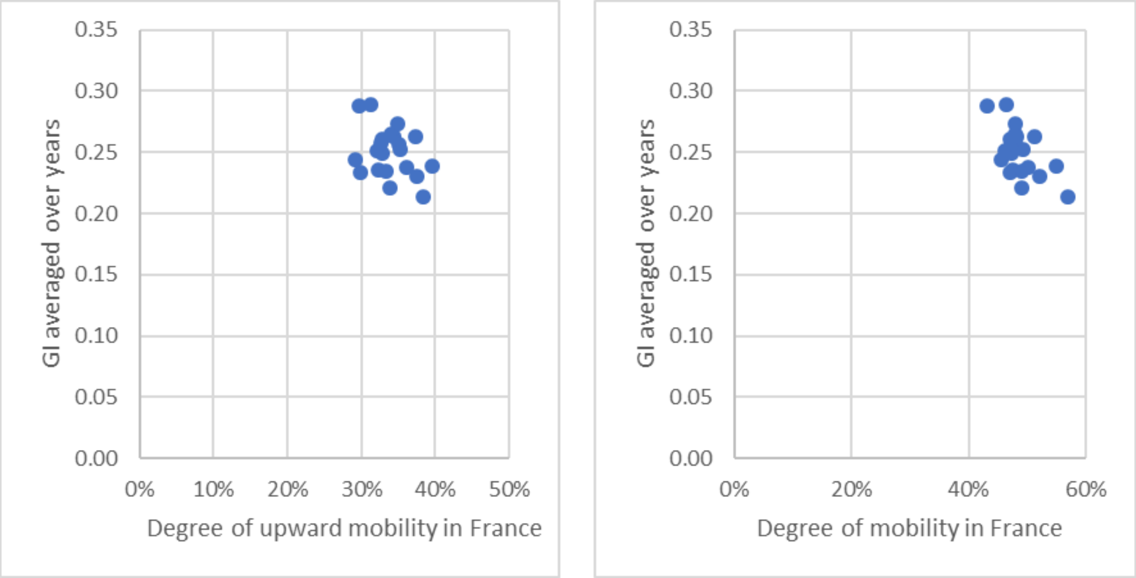


Figure 40: Upward and overall mobility and Gini index for France

3. Analysis of inequality of opportunity (IOp)

In this sub-section, we report and discuss the results obtained from estimations of the inequality of opportunity (IOp) at the national level, across NUTS2 regions, and by degree of urbanisation. As noted earlier, the term inequality of opportunity refers to the difference in individual economic outcomes that results from individuals' circumstances (i.e. factors they cannot control) assuming similar levels of effort (i.e. factors over which individuals have control).

The IOp approach here is based on the estimation of conditional income, where the economic outcome is again (gross) labour income and the explanatory variables are the individuals' circumstances. Specifically,

$$\tilde{y}_{it} = E[y_{it}|\boldsymbol{C}],$$

where y_{it} is labour income of individual i at time t and $\boldsymbol{C}$ is the matrix of circumstances beyond the control of the individual. Given that income is a continuous variable with inherent scale, we follow the OLS estimation approach with non-parametric methods by averaging over groups of individuals sharing the same circumstances, as proposed by Ferreira & Gignoux

(2011) to estimate $\tilde{y}_{it}$. Inequality of opportunity is computed using an absolute measure on $\tilde{y}_{it}$:

$$\theta_a = I(\tilde{y}_{it}),$$

where $I(.)$ denotes a common inequality measure. All variation in conditional income is solely due to circumstances and is therefore an absolute measure of the *level* of IOp. Ferreira and Gignoux (2011) use the mean logarithmic deviation to estimate this. Dividing θ_a by the measure $I(.)$ applied on actual labour income yields a relative measure of IOp:

$$\theta_r = \frac{\theta_a}{I(y_{it})}.$$

The latter is *relative* to overall income inequality. In order to further decompose IOp into its sources, we use the *Shapley decomposition*, whereby inequality measures for all possible permutations of circumstances are first estimated and then the average marginal effect for each circumstance on the measure of IOp is computed in order to estimate the relative importance of each circumstance (See Ferreira and Gignoux, 2014 for more details).

The foregoing analysis is based on Stata's user-written command iop (Juárez and Soloaga, 2014). The data on individuals' circumstances are extracted from two EU-SILC'S ad-hoc modules: the 2005 module on inter-generational transmission of poverty and the 2011 module on intergenerational transmission of disadvantages. The data contained in both these modules essentially refer to the socio-economic background of the respondent's family when aged around 14 years old. This type of information is highly relevant for the study of social mobility as it can capture the role of family background and home background, two crucial elements of child development and thus later life outcomes. The construction of the dataset underlying the analysis follows the same logic as that of the longitudinal dataset (described in Appendix B) except that there are no merges across different releases. The code with the process including the harmonization of variables between the two modules is available upon request by the authors.

The data in the two ad-hoc modules are included in EU-SILC's cross-sectional data, which, by construction, cannot be merged with the longitudinal data. As a result, we can only calculate individuals' IOp for the pair of years 2005 and 2011. Moreover, some of the circumstance

Co-funded by the Horizon 2020
Framework Programme of the European Union

variables, though similar in both models, are coded and/or labeled differently, requiring us to harmonize the variables in both modules to safeguard their correspondence. Our vector of circumstance variables includes the age and gender of the individual as well as a set of variables that capture the individual's family background, namely: education of father; education of mother; activity status of father; main occupation of father; activity status of mother; and the level of household financial situation.[22] We thus have 8 circumstance variables. In the Table below we describe the categories of the family background variables used in the IOp analysis.

We do not take into account the main occupation of the individual's mother because more than 40% of the observations are missing values that report to the fact that the mother is not working, most of which is already captured in the activity status of the mother and corresponds to fulfilment of domestic tasks and care responsibilities. For robustness, we checked pairwise correlations for all variables and there is no evidence of possible problems due to multicollinearity. We include age and gender in the model specification as control variables to capture variation in income levels resulting from different phases of individual's career life and gender-specific differences, some of which can admittedly include discriminatory processes.

[22] Education of both father and mother pertains to the highest level education attained. The main occupation is coded according to the ISCO-08 (COM) classification published by the International Labour Office where we add a code for those who have no occupation.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 29: Description of variables pertaining to family background

Variable	Code	Category
Education of father/mother	0	Could neither read nor write in any language
	1	Low level (pre-primary, primary education or lower secondary education)
	2	Medium level (upper secondary education and post-secondary non-tertiary education)
	3	High level (first stage of tertiary education and second stage of tertiary education)
Activity status of father/mother	1	Employed
	2	Self-employed (including family worker)
	3	Unemployed
	4	In retirement or in early retirement or had given up business
	5	Fulfilling domestic tasks and care responsibilities
Main occupation of father	0	Armed Forces Occupations
	1	Managers
	2	Professionals
	3	Technicians and Associate Professionals
	4	Technicians and Associate Professionals
	5	Services and Sales Workers
	6	Skilled Agricultural, Forestry and Fishery Workers
	7	Craft and Related Trades Workers
	8	Plant and Machine Operators and Assemblers
	9	Elementary Occupations
	10	No occupation
Financial situation of the household	1	Very bad
	2	Bad
	3	Moderately bad
	4	Moderately good
	5	Good
	6	Very good

We first analyze IOp results for Spain. Unfortunately, no information regarding gross labour income was reported for Spain in the year of 2005 (only net values reported), which means that we can only carry the analysis for 2011. We have 8 628 observations and provide a summary of the distribution of family background variables in the table below.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 30: Number of observations for each category of circumstances variable related to family background for Spain

Variable	Code	Number of observations	Percentage
Education of father	0	236	2.7%
	1	6876	79.7%
	2	652	7.6%
	3	864	10.0%
Education of mother	0	383	4.4%
	1	7335	85.0%
	2	495	5.7%
	3	415	4.8%
Activity status of father	1	6453	74.8%
	2	1960	22.7%
	3	29	0.3%
	4	119	1.4%
	5	67	0.8%
Activity status of mother	1	1564	18.1%
	2	589	6.8%
	3	13	0.2%
	4	15	0.2%
	5	6447	74.7%
Main occupation of father	0	134	1.6%
	1	548	6.4%
	2	471	5.5%
	3	811	9.4%
	4	564	6.5%
	5	811	9.4%
	6	1117	12.9%
	7	1719	19.9%
	8	1091	12.6%
	9	1147	13.3%
	10	215	2.5%
Financial situation of the household	1	211	2.4%
	2	677	7.8%
	3	1342	15.6%
	4	3571	41.4%
	5	2650	30.7%
	6	177	2.1%

For the regional analysis, we drop NUTS2 regions with less than 100 observations, which leads to the exclusion of Ceuta and Melilla. The following table summarizes the results obtained for 2011.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Table 31: IOp results for Spain in 2011 at the National level, by degree of urbanisation, and by NUTS2 regions

Geographical level		Number observations	iop	std error
National		8628	0.19	0.01
Degree of urbanisation	Large urban areas	4545	0.20	0.01
	Small urban areas	1776	0.18	0.02
	Rural areas	2307	0.18	0.02
Region	ES11	520	0.32	0.05
	ES12	344	0.31	0.06
	ES13	249	0.36	0.10
	ES21	490	0.14	0.03
	ES22	332	0.26	0.05
	ES23	304	0.18	0.04
	ES24	468	0.20	0.05
	ES30	983	0.21	0.03
	ES41	530	0.20	0.04
	ES42	447	0.26	0.07
	ES43	290	0.36	0.09
	ES51	1049	0.21	0.03
	ES52	727	0.19	0.03
	ES53	290	0.15	0.05
	ES61	815	0.21	0.03
	ES62	317	0.15	0.03
	ES70	383	0.23	0.05

The table displays the number of observations, the IOp relative measure (Ferreira and Gignoux, 2011) and the corresponding standard errors for the country-wide analysis and the analyses by degree of urbanisation and NUTS2 regions. All coefficients are statistically significant for a 1% significance level.

At the national level, we observe that 19% of the variation in gross labour income is due to circumstances uncontrollable by the individual. Inequality of opportunity is higher (0.2) in large urban areas compared to small urban areas and rural areas, being the same in the latter (0.18). At the regional level, IOp is very heterogeneous across regions, ranging from 0.14 in Basque Country to 0.36 in Extremadura. This means that in Extremadura 36% of the variation in gross labour income is due to individual's circumstances rather than effort, while in the Basque country individual's circumstances account only for 14% of the total variation in labour income. Aragón, Andalucía, Castilla y Léon, Cataluña and Comunidad de Madrid are the regions whose IOp is more on par with the IOp at the national level.

The following Figure shows the correlation between region population and IOp (to the left) and between the IOp and the GI (to the right).

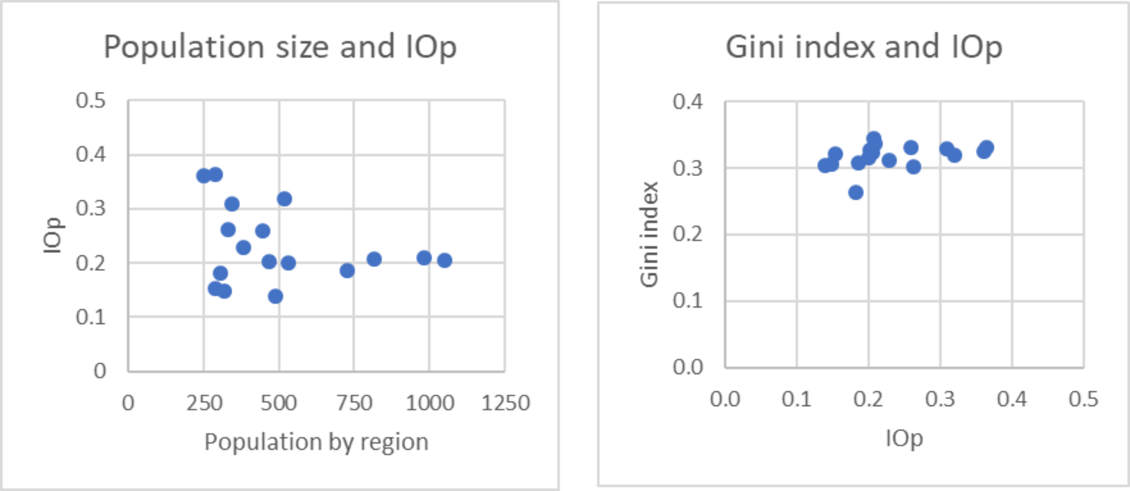


Figure 41: Correlation between region population size and IOp and between IOp and Gini index, in Spain, for 2011

The correlation between population size and IOp is negative (-0.28) but not statistically significant. Furthermore, the relationship does not appear to be linear. As for the relation between the GI and IOp there seems to be a slightly positive linear correlation between income inequality and the degree of inequality of opportunity, although the correlation coefficient is also not statistically significant (p-value is 0.14). Given that we observed a negative relationship between the GI and upward income mobility in the last section, this suggests a negative relationship between IOp and the level of upward income mobility.[23] Intuitively, this makes sense, as one would expect lower income inequality to be associated with lower inequality of opportunity and higher degree of upward mobility.

[23] It is impossible to draw a scatter plot because mobility indices refer to any pair of years over the entire period from 2005-2016, whereas IOp here is presented for a single year.

The following table displays the Shapley decomposition of IOp at the national level for the different circumstance variables considered. Age accounts for more than half of the IOp, but as mentioned earlier this is not per se a genuine source of inequality but rather a reflection of differences in individual's career stages. Gender accounts for 16% of the IOp relative measure and is followed by the main occupation of the father (10%), father's education (9%), household financial situation (6%), and mother's education (4%). The remaining variables account for very little of the IOp measure.

Table 32: Shapley Decomposition of IOp for Spain

Shapley decomposition	**2011**
Variable	Percentage of composition
age	51.57%
sex	16.46%
education of father	8.97%
education of mother	4.21%
activity status of father	0.42%
activity status of mother	2.20%
main occupation of father	10.16%
financial situation	6.01%
TOTAL	100.00%

We now discuss the results for France. In the 2005 ad-hoc module, the variable pertaining to financial situation only has missing values and is thus dropped from the analysis. The analysis uses 6 272 observations for 2005 and 6 810 for 2011. We provide a summary of the distribution of family background variables in the table below.

Table 33: Number of observations for each category of circumstances variable related to family background for France

		2005		**2011**	
Variable	Code	Observations	Percentage	Observations	Percentage
Education of father	0	282	4.5%	185	2.7%
	1	3884	61.9%	5140	75.5%
	2	1519	24.2%	634	9.3%
	3	587	9.4%	851	12.5%
Education of mother	0	310	4.9%	295	4.3%
	1	4480	71.4%	5164	75.8%
	2	1088	17.3%	694	10.2%
	3	394	6.3%	657	9.6%

		2005		2011	
Activity status of father	1	4749	75.7%	5464	80.2%
	2	1397	22.3%	1178	17.3%
	3	20	0.3%	24	0.4%
	4	26	0.4%	48	0.7%
	5	80	1.3%	96	1.4%
Activity status of mother	1	2450	39.1%	3236	47.5%
	2	416	6.6%	580	8.5%
	3	6	0.1%	11	0.2%
	4	4	0.1%	16	0.2%
	5	3396	54.1%	2967	43.6%
Main occupation of father	0	0	0.0%	104	1.5%
	1	731	11.7%	634	9.3%
	2	594	9.5%	582	8.5%
	3	488	7.8%	917	13.5%
	4	352	5.6%	538	7.9%
	5	192	3.1%	270	4.0%
	6	812	12.9%	684	10.0%
	7	1528	24.4%	1054	15.5%
	8	1074	17.1%	378	5.6%
	9	484	7.7%	1481	21.7%
	10	17	0.3%	168	2.5%
Financial situation of the household	1	-	-	190	2.8%
	2	-	-	488	7.2%
	3	-	-	1113	16.3%
	4	-	-	3002	44.1%
	5	-	-	1707	25.1%
	6	-	-	310	4.6%

The following table displays the IOp analysis for both 2005 and 2011.

Table 34: IOp in France for 2005 and 2011 at the National level, by degree of urbanisation, and by NUTS2 regions

Geographical level		**2005**			**2011**		
Level		obs	iop	std error	obs	iop	std error
National		6272	0.22	0.01	6810	0.21	0.01
Degree of urbanisation	Large urban areas	2979	0.23	0.02	2930	0.22	0.02
	Small urban areas	2312	0.22	0.02	2593	0.20	0.02
	Rural areas	981	0.13	0.02	1287	0.16	0.02
Region	FR10	1332	0.23	0.03	1125	0.22	0.03
	FR21	142	0.10	0.06	141	0.26	0.09
	FR22	203	0.24	0.06	229	0.17	0.06

Co-funded by the Horizon 2020
Framework Programme of the European Union

Geographical level		2005			2011		
	FR23	143	0.46	0.13	174	0.33	0.07
	FR24	301	0.18	0.05	305	0.22	0.06
	FR25	147	0.29	0.10	150	0.24	0.09
	FR26	166	0.22	0.07	191	0.18	0.08
	FR30	442	0.20	0.04	476	0.24	0.06
	FR41	245	0.19	0.06	339	0.24	0.06
	FR42	215	0.22	0.05	207	0.21	0.06
	FR43	159	0.24	0.06	171	0.26	0.09
	FR51	435	0.25	0.05	445	0.29	0.05
	FR52	296	0.26	0.05	391	0.27	0.05
	FR53	210	0.22	0.06	251	0.20	0.07
	FR61	277	0.26	0.06	395	0.19	0.05
	FR62	229	0.16	0.05	305	0.25	0.05
	FR71	570	0.23	0.04	628	0.22	0.04
	FR72	128	0.39	0.16	151	0.10	0.09
	FR81	167	0.23	0.07	238	0.19	0.07
	FR82	386	0.21	0.05	382	0.25	0.05

The level of inequality of opportunity in France was slightly lower in 2011 compared to 2005 (21% vs 22%), but still higher compared to Spain, meaning that a slightly higher proportion of income variability in France is due to individual circumstances such as family background. For both time periods, IOp is increasing with region population size, being lower in small urban areas compared to large urban areas, but pronouncedly lower in rural areas, particularly in 2005. The differences in IOp by degree of urbanisation are much higher in France compared to Spain.

As in Spain, there is a large heterogeneity in IOp across French regions, in both years. In 2005, IOp ranges from 0.10 in Champagne-Ardenne to 0.46 in Haute-Normandie. Although the number of observations is reduced for both regions, the latter IOp indicates that almost half of differences in income is due to individual circumstances. The overall heterogeneity is less pronounced in 2011 but is relatively scale invariant across regions between both years. Provence-Alpes-Côte d'Azur, Bourgogne, Poitou-Charentes show an IOp identical to that at the National level. The more populated regions, namely Île de France and Rhône-Alpes.

The following figures shows the correlation between region population and IOp (to the left) and between the IOp and the GI (to the right), for 2005 and 2011.

Co-funded by the Horizon 2020
Framework Programme of the European Union

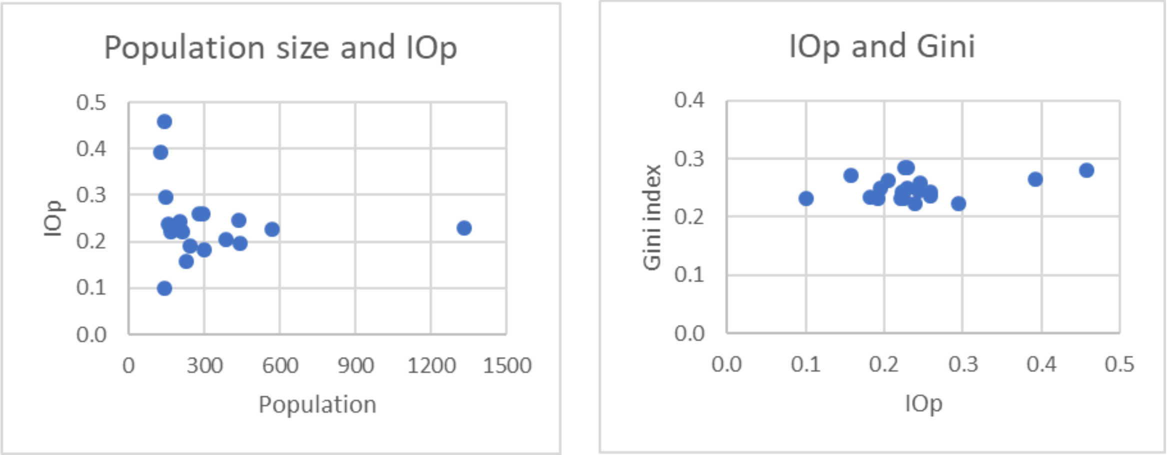


Figure 42: Correlation between region population size and IOp and between IOp and Gini index, In France, for 2005

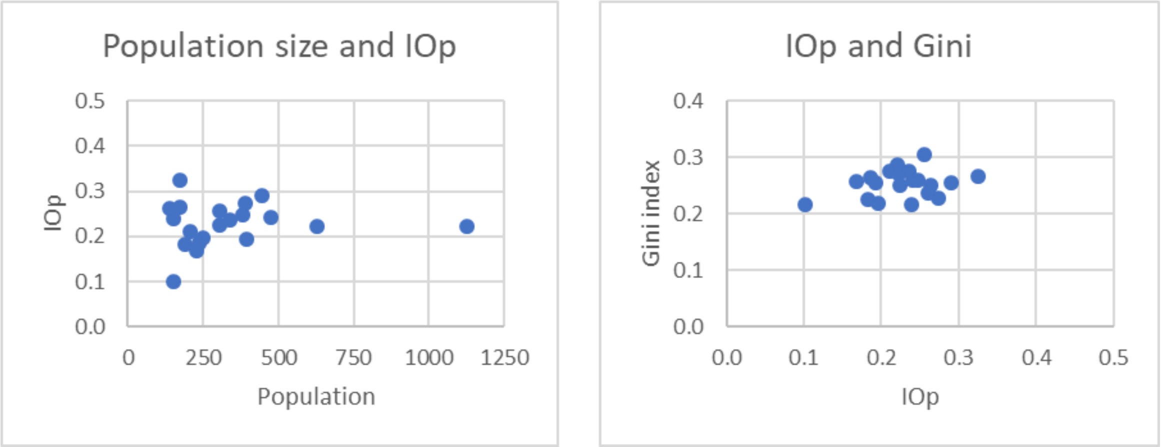


Figure 43: Correlation between region population size and IOp and between IOp and Gini index, in France, for 2011

In both 2005 and 2011, we observe no statistically relevant relationship between population size and IOp. On the other hand, and as was the case for Spain, there appears to be a positive relationship between the Gini index and IOp although the correlation coefficient for 2005 and 2011, 0.2 and 0.4 respectively, are not statistically significant.

Given the apparent positive relationship between income inequality and inequality of opportunity in France, together with the negative correlation between upward mobility and income inequality found in the previous subsection, we can again infer a negative relationship between IOp and income mobility. In other words, a higher equality of opportunity is related to higher income mobility, as expected.

The following table displays the Shapley decomposition of IOp at the National level for both 2005 and 2006. Age and gender account for the greatest part of the IOp, and are followed by father main occupation, father education, and mother education. Overall, aside from age and gender, there is an evident trend in the decomposition by circumstances if we compare both countries. The major difference lies in the importance of financial situation, which accounts for less than 1% of the IOp in France (only available for 2011).

Table 35: Shapley Decomposition of IOp for France in 2005 and 2011

Shapley decomposition	**2005**	**2011**
Variable	Percentage of composition	
age	29.02%	27.21%
sex	32.80%	38.82%
education of father	10.78%	13.15%
education of mother	8.23%	5.65%
activity status of father	0.39%	0.11%
activity status of mother	0.33%	0.39%
main occupation of father	18.45%	13.89%
financial situation	N/A	0.78%
TOTAL	100.00%	100.00%

4. Regression models of relative contribution of NUTS2 income deprivation to individual labour-income

In this sub-section, we report and discuss the results obtained from the microeconometric regression models of in individual income levels. The models measure the relative contribution of individual characteristics and NUTS2-level income deprivation to individual income levels. Given the short time span (i.e. maximum of four years for each individual) and the rather aggregate nature of the geographical units available (i.e. NUTS2 regions and a simple indicator of degree of urbanisation), it is not possible to implement more advanced econometric techniques. Nevertheless, we can consider the main panel data estimators, namely: pooled OLS, fixed-effects (FE), random-effects (RE), and the mundlak estimator (Mundlak, 1978) for correlated random-effects (CRE).

Of the three EU-SILC countries considered so far it was only possible to measure the effect of NUTS2-level income deprivation on individual income level for Spain and for Finland because there is no regional NUTS2 data for at-risk-of-poverty rates for France. In addition, changes in NUTS2 regions in Finland only permit using data from 2008 onwards. The table below summarises the main findings for the two countries. In both cases, taking account of the panel data component by using random-effects (RE) and fixed-effects (FE) estimators leads to a reduction in the magnitude of the coefficient for NUTS2-level income deprivation. Moreover, accounting for annual variation reduces considerably the level of statistical significance. In the case of Finland, including controls for years results in no effect from regional income deprivation on individual income, while for Spain there is still a significant effect for the pooled OLS and the RE models. According to these two estimators, an increase of 1 point in the regional at-risk-of-poverty rate is associated with a reduction in individual income level of about 0.22% and 0.10% respectively. In other words, at this large spatial scale the relationship seems to be small, which is in line with the national analyses implemented in the following sections.

Table 36: EU-SILC regression models for Spain (top) and Finland (bottom)

Spain	**Pooled OLS**	**RE**	**FE**	**Pooled OLS**	**RE**	**FE**
NUTS2-level AROP rate	-0.438***	-0.188***	-0.093**	-0.224***	-0.098**	-0.044
female	-0.282***	-0.289***	-0.267***	-0.281***	-0.290***	-0.273***
age	0.015***	0.015***	0.014***	0.015***	0.015***	0.013***
constant	9.525***	8.835***	9.085***	9.470***	8.794***	9.086***
Controls for education	Yes	Yes	Yes	Yes	Yes	Yes
Controls for occupation	Yes	Yes	Yes	Yes	Yes	Yes
Controls for degree of urbanisation	Yes	Yes	Yes	Yes	Yes	Yes
Controls for NUTS2	Yes	Yes	Yes	Yes	Yes	Yes
Controls for year	NO	NO	NO	Yes	Yes	Yes
Observations	103263	103263	103263	103263	103263	103263
Adj R2	0.42			0.425		
r2_overall	0.42	0.417	0.378	0.425	0.422	0.383
r2_between		0.425	0.383		0.429	0.386
r2_within		0.077	0.082		0.09	0.097
Finland	**Pooled OLS**	**RE**	**FE**	**Pooled OLS**	**RE**	**FE**
NUTS2-level AROP rate	-1.764***	-1.050***	-1.083***	-0.55	-0.189	-0.163
female	-0.255***	-0.276***	-0.306***	-0.254***	-0.275***	-0.305***
age	0.015***	0.018***	0.020***	0.015***	0.018***	0.017***
constant	9.318***	8.989***	8.885***	9.083***	8.826***	8.814***
Controls for education	Yes	Yes	Yes	Yes	Yes	Yes
Controls for occupation	Yes	Yes	Yes	Yes	Yes	Yes
Controls for degree of urbanisation	Yes	Yes	Yes	Yes	Yes	Yes
Controls for NUTS2	Yes	Yes	Yes	Yes	Yes	Yes
Controls for year	NO	NO	NO	Yes	Yes	Yes
Observations	27044	27044	27044	27044	27044	27044
Adj R2	0.353			0.353		
r2_overall		0.347	0.314		0.348	0.315
r2_between		0.352	0.320		0.350	0.316
r2_within		0.106	0.115		0.113	0.122

*, **, *** denote statistical significance p<0.10, p<0.05, p<0.01 respectively.

Co-funded by the Horizon 2020
Framework Programme of the European Union

Appendix E

Description of data sources and spatial units used in the computation of bespoke neighbourhood income deprivation in task 5.2 for England and Scotland

As documented in report for Deliverable Task 5.2, the best direct measures of income available for small geographical areas were the small area estimates of household income, which are produced by the Office for National Statistics (ONS) for MSOAs in England and the Scottish Neighbourhood Statistics for DZs (i.e. LSOAs) in Scotland. As noted therein, these small area income statistics consist of model-based small-area estimates and thus are not actual measures of household income.

In the case of England, the income deprivation indicator is calculated using a model-based method to produce estimates of household income using a combination of survey data from the Family Resources Survey and previously published data from the 2011 Census and a number of administrative data sources. The estimates are available at the level of middle layer super output area (MSOAs) in England for 2011/12. MSOAs have a mean population of 7,200 and a minimum population of 5,000. They are built from groups of LSOAs and constrained by local authority boundaries. The indicator consists of MSOA level estimates of the proportion and count of households below 60% of the UK median income after housing costs (AHC) and before housing costs (BHC) for 2013/14. The analysis uses the before housing cost indicator in order to make it more comparable to the Scottish case, for which there is no after housing cost at-risk-of-poverty rates indicator.

Similarly, the income deprivation indicator for Scotland consists of small area model-based income estimates for the average household gross weekly income, referring to 2014 and Data Zones[24] (DZs, and correspond to the Lower Super Output Areas (LSOAs)), which covers total

[24] Data zones are the main small-area statistical geography in Scotland and consist of groupings of Census Output Areas (OAs). DZs correspond to the Lower Super Output Areas (LSOAs). DZs have populations of between 500 and 1,000 household residents, and there were 6,500 2001 DZs and 6,976 2011 DZs in Scotland for each census period respectively.

Co-funded by the Horizon 2020
Framework Programme of the European Union

income received by all adult members of a household, including welfare benefits, tax credits and housing benefit. The estimates reflect total income before any deductions are taken off for income tax, national insurance contributions and council tax etc. The indicator used is the count of households with gross household weekly income below 60% of the median national (Scottish) income. From this count it is possible to obtain the at-risk-of-poverty (AROP) rates. The information available refers only to household income before taking account of housing cost. 4 Finally, it is important to note that the Scottish islands have been excluded from the EquiPop analysis due to issues relating to the very sparse nature of population settlements in these remote areas (and also in the Highlands).

In summary:

- England - Count and ratio of household with mean gross income below 60% of the UK median income (before housing costs). MSOA-level data for 2013/14.
- Scotland - Count and ratio of household with mean gross income below 60% of the UK median income (before housing costs). LSOA (=DZ) level data for 2014.

Using the EquiPop software we calculated in Task 5.2 the proportion of individuals with a low income at different spatial scales. Starting with the proportion of people with a low income among the nearest 200 people, up to the percentage of people with a low income among the nearest 51,200 people. By increasing the scale, the contextual variable of each grid cell (which is a proxy for a residential location) measures poverty for a larger population, and by definition also a larger geography. The data for the UK starts with relatively large building blocks, which already contain 3,248 households on average (e.g. MSOAs for England), we could not meaningfully calculate the poverty rate for k=200 and k=1,600. As the underlying geographies already contain more households, there would not be any differences between measures calculated for the lowest spatial scales.

Co-funded by the Horizon 2020
Framework Programme of the European Union