I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15937

The Transformation of Public Policy Analysis in Times of Crisis – A Microsimulation-Nowcasting Method Using Big Data

Cathal O'Donoghue
Denisa M. Sologon

FEBRUARY 2023

IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15937

The Transformation of Public Policy Analysis in Times of Crisis – A Microsimulation-Nowcasting Method Using Big Data

Cathal O'Donoghue
National University of Ireland and IZA

Denisa M. Sologon
Luxembourg Institute of Socio-Economic Research and IZA

FEBRUARY 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Transformation of Public Policy Analysis in Times of Crisis – A Microsimulation-Nowcasting Method Using Big Data

The urgency of the two crises, especially the COVID-19 pandemic, revealed the inadequacy of traditional statistical datasets and models to provide a timely support to the decision-making process in times of volatility. Drawing upon advances in data analytics for public policy and the increasing availability of real-time data, we develop and evaluate a method for real-time policy evaluations of tax and social protection policies. Our method goes beyond the state-of-the-art by implementing an aligned or calibrated microsimulation approach to generate a counterfactual income distribution as a function of more timely external data than the underlying income survey. We evaluate the simulation performance between our approach and the transition matrix approach by undertaking a nowcast for a historical crisis, judging against an actual change and each other. Nowcasting emerges as a useful methodology for examining up-to-date statistics on labour force participation, income distribution, prices, and income inequality. We find significant differences between approaches when the calibration involves structural heterogenous changes. The model replicates the changes in income distribution over one year; over the longer term, the model is able to capture the trend, but the precision of the levels weakens the further we get from the estimation year.

JEL Classification: I31, I38, C54

Keywords: big data, policy analysis, nowcasting, microsimulation, COVID-19

Corresponding author:
Cathal O'Donoghue
National University of Ireland
University Road
Galway
Ireland

E-mail: cathal.odonoghue@nuigalway.ie

The Transformation of Public Policy Analysis in Times of Crisis – A Microsimulation-Nowcasting Method using Big Data

1. Introduction

The shock triggered by the COVID-19 crisis has altered the *policy-modelling-data nexus*, a change that had already begun after 2008/2009 financial crisis. The asymmetric nature of both crises, the periods of increased heterogeneity and volatility made it difficult for policy makers to understand and assess in real time the socio-economic impact of these crises. This revealed an inadequacy of traditional statistical datasets and models to provide a timely support to the decision-making process in times of economic volatility.

The needs of policymakers for timely quality data and analysis to enable faster and better decision making, have accelerated the efforts to obtain timelier, richer and more granular data for policy analysis; "big data" (Tissout, 2017; Nymand-Andersen, 2015).[1] This evolution is not new, but part of the two broad and interdependent data science trends: the gradual expansion of new types of data and statistical methods that can be used to inform public policy and the development of new tools in data analysis (Engler, 2020). Undoubtedly, the expansion of new types of data and statistical methods, the compilation of rich data resources linked and analysed with "big data" tools are changing the landscape of policy modelling and its applicability in timely decision-making, providing unprecedented opportunity for public policy analysis (Jarmin and O'Hara, 2016).

The COVID-19 shock created exceptional demands that forced public sector systems (such as those delivering social protection measures) to change; it created challenges, but also opportunities. New policies had to be put in place swiftly to deal with the urgent social and economic effects of the pandemic. This in turn challenged the systems to provide timely data to evaluate these policies, creating a surge in transparency, volume, velocity, variety and veracity in data delivery (March and Marcus, 2021). Administrative systems were put in place during the early stages of the crisis to allow for close to real-time reporting by social security, resulting in better-organized data available in the public domain, with recent releases on which modellers can build. The ABS Single Touch Payroll data in Australia that allowed the Taxation Office to receive payroll information from the accounting software each time the business run its payroll (Li et al., 2022). Similarly, in Ireland, the COVID social security system and the wage subsidies that went to the tax systems reported regularly, providing information within a week or two (O'Donoghue et al., 2020).

The expansion in data availability and timeliness enabled methodological advancements to facilitate close to real time decision making (Brewer and Tasseva, 2020; Bronka et al., 2020; O'Donoghue et al., 2020; Li et al., 2022; O'Donoghue et al., 2022; Sologon et al., 2022), particularly in the area of microsimulation modelling which is a long-established analytical tool for understanding the impact of public policy on different population groups (O'Donoghue, 2014); microsimulation models provided data analytics using "big data" long before the emergence of the "big data" revolution.

Rapid economic changes put pressure on public policy in terms of its ability to continue to meet its objectives and in terms of the potential cost of the system. While household income

[1] Defined as "data generated as a consequence of government, business, or citizen activities" and characterized by volume, velocity, variety and veracity (Jarmin and O'Hara, 2016).

survey data have carefully designed purposes, they fall short on volume, velocity and veracity (Jarmin and O'Hara, 2016). As survey data is produced with a time lag of 2-3 years, in times of volatility, the data required to do appropriate policy analysis can date relatively quickly. These data lags are problematic for policymaking, particularly in times of economic volatility. In order to support policymaking in times of crisis, it is needed to adjust the distributional characteristics of a household income survey dataset to account for changes between data collection and analysis a process known as nowcasting (O'Donoghue and Loughrey, 2014). Nowcasting or static ageing (Immervoll, 2005) involving price adjustments and policy parameter adjustments have had widespread use. Nowcasting in a time of large economic changes however requires adjustment to the underlying survey reflecting large changes in employment.

The pace of change during the COVID crisis has forced economic model builders to develop new ways to undertake economic impact assessment in order to make models more relevant with the timely outputs using nowcasting methods. Nowcasting changes to the labour market projects historical data to the present by calibrating with close-to-real time information, provided by administrative channels or "big data". Complementing survey data with "big data" has enabled new analytical tools to be built, thereby enhancing the uses of microsimulation in times of crisis. The use of microsimulation has become extensive during the COVID crisis to nowcast the distributional implications of labour market and social protection policy measures to manage the crisis (Brewer and Gardiner, 2020; Bronka et al., 2020; Figari and Fiorio, 2020; Li et al., 2022; Lustig et al., 2021; O'Donoghue et al., 2020; Sologon et al. 2022).[2] This combination of administrative and survey provides added-value to "big data" in the sphere of social protection as administrative data typically does not allow for detailed contextual analysis at the level of the household.

A variety of different nowcasting approaches have been developed. The papers based on the pan-European tax-benefit microsimulation model EUROMOD (Brewer and Gardiner, 2020; Bronka et al., 2020; and Figari and Fiorio, 2020) and Lustig et al., (2021) used transition matrix based methods to simulate exits from employment, to calibrate the changes observed in administrative data, but not yet available in household survey data. Our paper goes beyond the state-of-the-art by proposing an aligned or calibrated dynamic microsimulation approach to generate a counterfactual income distribution as a function of more timely external data than the underlying income survey (Li and O'Donoghue, 2013). The methodology allows for greater heterogeneity and can be used at different stages of the business cycle. We will compare the simulation performance between our approach and the transition matrix approach used by many other papers.

In order to evaluate the simulation performance of the method proposed, we undertake a nowcast for a historical crisis, evaluating each method against an actual change and against each other. As a test, we select the previous global crisis, the financial crisis or great recession. According to the Eurostat Labour Force Survey, Ireland had the largest loss of employment in the financial crisis, losing 14% of its employment between 2007 and 2011 (Savage et al., 2019). We will then apply the approach to the ongoing COVID crisis nowcasting over various waves of the crisis to autumn 2021. Ireland was also significantly affected with the second highest fall in employment during the first wave within the EU. The paper defines the nowcasting

[2] They are also used when one wishes to make projections of the income distribution from one point to another (See Brewer et al., 2013). This may also be required where one is assessing the impact of macro-economic shocks or policy changes on the distribution of income (Navicke et al., 2014).

methodology in a methodology annex, evaluates its simulation performance during the Financial Crisis and applies in the context of the COVID-19 crisis.

2. Methods and Data

The objective of our approach is to understand the impact of close to real time labour market, policy and price changes on the distribution of welfare. Nowcasting combines changes to the three components of disposable income:

- Income indexation – the change in the level of income resulting from changes to average wages;
- Tax-Benefit parametric and structural changes;
- Adjustments for changes in the population composition with respect to the demographics and with respect to the prevalence of income sources.

Income indexation (Immervoll, 2005) and tax-benefit updating (O'Donoghue et al., 2018) is undertaken using standard respectively uprating and microsimulation methods.

Nowcasting labour market change

Building upon the macro-economic literature (Giannone et al., 2008), there is a growing literature on the use of nowcasting for producing more timely estimates of inequality (O'Donoghue and Loughrey, 2014) including the modelling of poverty incidence (Álvarez et al., 2014). In this paper, however, we are interested in understanding the components that affect different parts of the income distribution and to understand the process by which asymmetric shocks impact different households in different ways.

Much of the literature relies on combining more timely labour force data, which has limited or non-existent income data, with income survey information. The EUROMOD nowcasted income distributions (Leventi et al., 2014; Navicke et al. 2014; Almeida et al., 2021) have had the highest visibility. They use the European Labour Force Survey to draw employment rates by demographic factors like age, gender, and education level, in order to simulate proportional changes in industry-specific employment rates. This information is then combined with the EUROMOD tax-benefit model to evaluate the policy outcomes. A number of papers have utilised EUROMOD in the COVID crisis to provide up-to-date distributional implications of the market and policy responses to the crisis (Figari and Fiorio, 2020; Brewer and Gardiner, 2020; Beirne et al., 2020; Brewer and Tasseva, 2021). Lusting et al. (2020) utilised a similar approach randomly allocating a change within a demographic group.

In other words for population sub-group i, with employment rate e_i and random number u, an individual is simulated to be in employment if

$$u < e_i$$

Nowcasting, the target employment probability moves from e_i to e_i^* and the employment rate is recalculated:

$$u < e_i^*$$

Addabbo et al., (2016) extended the parametric perspective of EUROMOD by modelling employment transitions using estimated probit equations: from unemployed and inactive to employed (with/out reduced hours), and from employed to inactive and unemployed. They also modelled the participation in the wage guarantee scheme. Employment income for those simulated to enter the labour market are then imputed using a Heckman selection model. In the case of an employment transition, the probability of employment for sub-group i depends in addition on additional demographic characteristics Z, contained in the probit model, $e(Z)_i$

$$u < e(Z)_i$$

Nowcasting they use "imputed probabilities to obtain probability thresholds according to the relative change in employment status within the strata".

$$u < \frac{e_i^*}{e_i} e(Z)_i$$

Carta (2019) took another approach; instead of taking the labour force status from household surveys, she imputes labour incomes into the Labour Force Survey. In doing that she draws upon recent labour distributions and adds modelled income utilising Mincerian wage equations. In other words, the paper does not simulate the nowcasted employment rate e_i^*, but the wage rate. While this improves knowledge about the heterogeneity of employment such as within household correlations and correlations with demographic factors, the paper has a different focus considering only wage income rather than all market incomes and consequential disposable incomes.

Li et al. (2021) take a semi-parametric perspective, drawing upon the methodology of (DiNardo et al., 1996).[3] They use the monthly Australian Longitudinal Labour Force Survey (LLFS) data, together with administrative payroll data. They utilise changes in the LLFS to reweight the labour market characteristics at the intensive and extensive margins of the labour market in the income survey along the lines of demographics, employment and household characteristics.

While convenient in multi-country simulations, the use of Monte-Carlo simulations based upon cell-specific probabilities may ignore some of the important heterogeneity exhibited in a crisis. For example, family status may be an important driver. Similarly, the situation of partners within a family may be correlated as identified in Carta (2019). Carta (2019) avoids issues associated with intra-household variations or sectoral biases by using recent labour force survey data. However, ignoring the impact of other non-labour force characteristics and, in particular, the impact of public policy as an insulating mechanism is an issue.

Dynamic ageing (see Li and O'Donoghue, 2013) involves the estimation of a system of econometric equations, which are in turn used to simulate changes in the population. As noted above, we refer to dynamic in the sense of a dynamic change to the distribution rather than the modelling of individual income dynamics. The approach starts with a sample of households whose underlying characteristics, are held constant. The sample's characteristics are altered using a dynamic simulation mechanism to generate various distributions, representing expected characteristics in either the future or the present (in the case of nowcasting). This method entails estimating a system of equations that replicate the distribution of incomes by its sources. Dynamic ageing typically models at the individual level handling inter-dependencies between

[3] Kump, and Navicke (2014) also use a reweighting approach to incorporated demographic changes.

individuals in a conditional way rather than jointly, with the latter imposing very significant computational requirements (Pudney, 1992).

Reweighting or semi-parametric approaches are strong in the sense that they avoid distributional assumptions. However, like other static-reweighting procedures they rely on the existence of sufficient sample sizes (Klevmarken, 1997). Increasing the number of dimensions in the analysis increases the risk of relying on cell weights with small numbers. Immervoll et al. (2005) identify a number of other issues:

- controlling for changes in aggregate group sizes may not improve the match for distributional patterns if much of the changes occur within the group;
- applying aging techniques in a mechanical manner across different time periods is not recommended since it can lead to inaccuracies. Structural changes in the population and tax-benefit system can greatly impact the suitability of a set of alignments, so they should not be done without considering these changes;
- adding information to a dataset by adjusting the statistical weights can impact the accuracy of the original weights, especially if multiple dimensions are used in the re-weighting process. The greater the number of dimensions, the greater the risk of skewing the data.

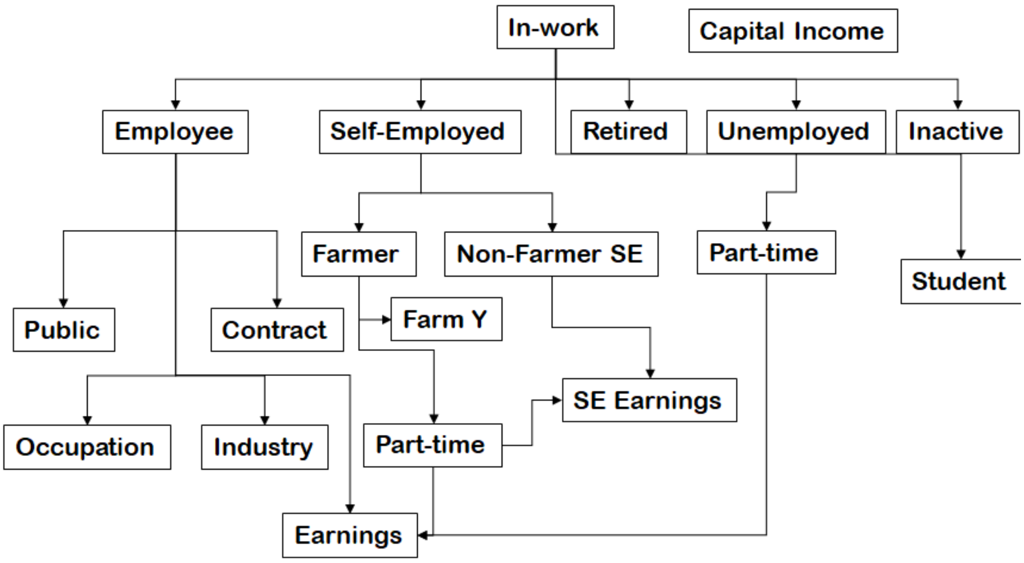
Utilising the DiNardo et al. (1996) semi-parametric approach has a higher data requirement than calibrated approaches, requiring access to micro data to generate counter-factual weights. Working at close to real time this may not be feasible. As a result, in this paper, we focus on calibrated approaches.

In the case where reweighting or semi-parametric approaches prove unfeasible, a parametric approach may be a more pragmatic approach. In this paper we consider in more detail a parametric approach akin to the Addabbo et al. (2016) methodology, but in addition drawing upon the alignment aspects of the dynamic microsimulation literature, dynamic ageing. We apply this approach for looking at the impact on the COVID crisis in Ireland (O'Donoghue et al., 2020). In this paper, we discuss in detail the methodology and we validate it using a short to medium term horizon.

To account for macro-economic changes, we opted for a dynamic aging mechanism for our income distribution data, due to the significant and fast-paced changes in the economy structure.

The precise nature of the methodology is described in the methodology annex. Fundamentally however the analysis involves producing a system of equations that describe the distributional characteristics of the labour market, known as an Income Generation Model (IGM), described in Figure 1. We estimate these models and save the parameter estimates and the residuals. These will in turn be used in the simulations, calibrations and alignments. A fundamental difference with the systems of equations that are used in dynamic microsimulation modelling is that the IGM only considers an error component structure that incorporates unobserved cross-section distributional and ignores inter-temporal transitions required in a dynamic model.

Figure 1. Income Generation Model



The methodology appendix describes how the estimated system of equations can be aligned to external calibration data derived from administrative "big" data.

Calibration data to index income growth

To utilize the dynamic aging methodology, multiple data sources are necessary. These include:
- calibration data to index income growth;
- calibration data to align the labour market;
- microdata for estimatimations/simulations.

Calibration totals on which to index income growth are taken primarily from the Irish Central Statistics Office, quarterly Earnings Hours and Employment Costs Survey, which is available at a lag of 1 quarter.

The basis of our analysis is a series of calibration control totals that show the shifts in the macro-economic conditions in Ireland from 2007-2009, which saw the greatest impact of the financial crisis, and from 2020-2021, particularly regarding the labor market's structure. Reflecting the structure of the labour market, we utilise control totals for the following variables
- In-Work by age group by gender
- Employee by gender
- Retired by gender
- Unemployed by gender
- Occupation Group (9 Category SOCEC) by gender
- Industrial Group (8 Category) by gender.

To comprehend the effects of alterations in the labor market, earnings, and policy decisions, data with ample detail is required. The EU-SILC, which has been gathered in Ireland since 2003, is an ideal dataset that serves this purpose. It replaces the previous European Community Household Panel Survey and collects data on earnings, labor market characteristics, demographics, and living conditions. This information is used to analyze poverty, inequality, and deprivation. We use the 2007 dataset to simulate the financial crisis.

The EU-SILC data is collected nationally and harmonized for Eurostat, which then processes it and provides it to researchers as a harmonized User Database (UDB). For modeling the income distribution, the Irish component of EU-SILC (UDB) is used. The data is in gross form, without taxes and contributions, and combines survey and register data, with 80% of respondents consenting to using their national social security number for accessing administrative data on benefit entitlements (Callan et al., 2010).

A national weighting approach that takes into account factors such as gender, age group, region, and household composition is utilized. This method is based on a blend of population projections from the Census and the Quarterly National Household Survey (Callan et al., 2010). However, it is important to note that while the weights depict the population structure, they do not accurately reflect either the social transfer recipients or the taxable income distribution. Callan et al. (2010) suggested using external data to refine the representativeness in these areas, but for the purpose of this paper, which is to compare the difference using the EU-SILC definition of income and its weights, this adjustment has not been made.

Challenges exist in using EU-SILC for microsimulation modelling.

- Parental and partner ID variables are available, allowing for creation of most within-household units of analysis needed for tax-benefit systems. However, the data falls short when knowledge of inter-household units of analysis, like for higher education grants, is required.
- The EU-SILC faces a challenge in that the income variables are based on the previous year while personal characteristics are based on the time of the interview, leading to potential inconsistencies. For example, a person may have been unemployed during the interview year but still show employment income in the data. Ireland has a slightly different definition, with an "income reference period" spanning two tax years and ending on the date of the interview, with approximately 25% of the sample collected each quarter.
- Both tax-benefit models and EU-SILC aim to measure household disposable income, and the EU-SILC generally has the necessary variables. However, there are some missing variables such as capital gains, wealth, and property values. This is a common issue in income surveys, and tax-benefit microsimulation models often use a definition of disposable income that does not include taxes based on these factors. It is reasonable for an EU-SILC-based model to adopt a similar approach.

Microsimulation modelling faces a challenge as some variables are not easily assignable to the proper unit of analysis. For instance, some income variables such as capital income and family benefits, recorded only at the household level, may be assigned to the head of household. This may lead to overestimation of taxes in a progressive tax system if the income was received by someone else in the household, leading to bias in results.

This aggregation of benefits into limited categories can lead to a loss of information, leading to an underestimation of the amount of benefits received by households. This can be particularly problematic in the case of social exclusion and housing benefits, as these benefits are often means-tested, meaning that the actual level of support provided can vary significantly from household to household.

We would like to use information about benefit receipt to model the level of social insurance benefits, since we don't have information about the contributory conditions for these benefits. Adequate information is present to model most social assistance and family benefits. Callan et al. (2010) have access to a special research edition of the 2008 EU-SILC, which does not have the aggregation problems.

However, even with fully modelled instruments, the knowledge of benefit presence and value is still important to model benefit take-up as some models assume 100% take-up.

Challenges also exist in using EU-SILC data to infer mis-calculation of taxes and social insurance contributions. Ideally, separate data on taxes and contributions would be available at the most relevant unit of analysis, but in EU-SILC they are reported in a single variable and at the household level. This limitation is common to many income datasets, but it still presents a challenge in making accurate inferences.

The EU-SILC contains some useful expenditures for tax-benefit modelling, such as mortgage interests and pension contributions. However, it lacks information on deductible expenses like medical insurance. It also does not have information on the value of residential properties, which is necessary to model the local property tax in Ireland. Validation of the tax-benefit model can be found in O'Donoghue et al. (2013).

3. Results

In theory, a regression-based nowcasting approach should enable greater heterogeneity to be incorporated relative to a transition probability approach. In Table 1 we report the characteristics of simulated people in work in 2009 off a base 2008 population with the actual characteristics for 2009. As both methods calibrate to actual employment levels, the aggregate employment rates are the same.

Table 1. Characteristics of In-Work Regression for 2009 simulated off 2008 Population

	Univ.	Upper -sec.	No. children 0-4	No. children 5-13	No. children 14-20	Married	Peri-Urban	Rural
Male								
Monte Carlo	0.31	0.37**	0.181	0.397	0.484**	0.553**	0.283	0.352
IGM	0.316	0.368**	0.182	0.399	0.472**	0.555**	0.281	0.356
Actual	0.312	0.311	0.182	0.455	0.351	0.610	0.275	0.370
Female Couple								
Monte Carlo	0.447	0.337	0.223	0.455**	0.412**	0.865	0.28	0.366
IGM	0.463	0.337	0.233	0.442**	0.366	0.852**	0.273	0.369
Actual	0.44	0.352	0.211	0.533	0.344	0.879	0.277	0.365
Female Single								
Monte Carlo	0.411	0.396	0.053	0.227	0.566**	0.013	0.282	0.229
IGM	0.43	0.387	0.06	0.219	0.545**	0.012	0.278	0.233
Actual	0.406	0.413	0.07	0.226	0.363	0.019	0.308	0.255

Note: Monte Carlo refers to the transition approach; IGM refers to the dynamic microsimulation approach and Actual refers to SILC-2009 (2008 incomes).

The table reports where the simulated share of a particular characteristic for workers is different to the actual data. As in the case of Navicke et al. (2014), the transition probability or Monte Carlo analysis uses transition rates differentiated by age and gender. To consider the implications for differential heterogeneity, we look at the impact of incorporating a wider set of explanatory variables. For males, the share of upper secondary educated is significantly different from the actual situation, while the share of university educated is similar. As a result, the models over simulate the loss of employment by those with a lower secondary education. There are also slight differences in the number of children aged 14+ and the share of married. We note that where there is a statistically significant difference for the Monte Carlo simulation, there is also a difference in the same difference in the same direction for the Income Generation Model (IGM) simulation.

We separate females by partnership status given different labour market patterns. For females in a couple, there are few differences with the actual data, except for the numbers of children aged 5-13, who are under-represented in both models. Older children are over-represented in the Monte Carlo simulation, while married women are slightly under-represented in the IGM. The only statistical difference for single women is for older children.

On the face of it, there are few differences between both simulations. Nevertheless there are some small statistically significant differences in terms of heterogeneity with the actual data.

Table 2. F- Test for the In-work transition, IGM versus Monte Carlo

Var #	Gender	Male		Female		Male		Female	
		Beta	S.E.	Beta	S.E.	Beta	S.E.	Beta	S.E.
1	University	1.2856***	0.1051	1.7789***	0.0955	1.2856***	0.1051	1.7789***	0.0955
2	Upper Secondary Education	1.1107***	0.0878	1.0486***	0.0809	1.1107***	0.0878	1.0486***	0.0809
3	Numer of Children Aged 0-3	0.1086	0.1235	-0.677***	0.0912	0.1086	0.1235	-0.677***	0.0912
4	Number of Children aged 4-11	-0.1594**	0.0626	-0.4843***	0.0518	-0.1594**	0.0626	-0.4843***	0.0518
5	Number of Children aged 12-15	-0.2285***	0.0779	-0.4317***	0.0685	-0.2285***	0.0779	-0.4317***	0.0685
6	Married	0.653***	0.0959	-0.366***	0.0860	0.653***	0.0959	-0.366***	0.0860
7	Age	0.3168***	0.0103	0.3151***	0.0120	0.3168***	0.0103	0.3151***	0.0120
8	Age Squared	-0.0038***	0.0001	-0.004***	0.0001	-0.0038***	0.0001	-0.004***	0.0001
9	Per-Urban	-0.0884	0.0918	0.0237	0.0877	-0.0884	0.0918	0.0237	0.0877
10	Rural	0.4898***	0.0890	-0.0793	0.0838	0.4898***	0.0890	-0.0793	0.0838
11	Year == 2009				0.2283	-0.5203	0.3162	-1.0083***	0.3457
12	Year == 2009 x University					0.0949	0.1430	-0.122	0.1330
13	Year == 2009 x Upper Secondary Education					-0.2146*	0.1241	0.0299	0.1163
14	Year == 2009 x Numer of Children Aged 0-3					0.1195	0.1728	0.0987	0.1248
15	Year == 2009 x Number of Children aged 4-11					0.0092	0.0859	-0.0044	0.0712
16	Year == 2009 x Number of Children aged 12-15					0.0013	0.1127	0.0516	0.1000
17	Year == 2009 x Married					-0.052	0.1342	0.2291*	0.1198
18	Year == 2009 x Age					0.0107	0.0150	0.0259	0.0180
19	Year == 2009 x Age Squared					-0.0001	0.0002	-0.0002	0.0002
20	Year == 2009 x Peri-Urban					-0.1125	0.1292	-0.0669	0.1235
21	Year == 2009 x Rural					-0.1392	0.1247	0.0111	0.1182
22	Constant	-5.7354***	0.2159	-5.2276***		-5.7354***	0.2159	-5.2276***	0.2283
	Pseudo R2	0.4195		0.4052		0.4209		0.4021	
	Number Obs	6079		6472		12208		12984	
	Chow Test								
	Variables Tested	1 2 3 4 5 6 9 10				11 12 13 14 15 16 17 18 19 20 21			
	Likelihood	337.98		588.93		51.41		47.21	
	p-value	0		0		0		0	

In Table 2, we run several Chow Tests to understand differences between models and differences over time. In the first model we estimate the relationship between in-work and various explanatory factors for males and females. We run an F-test to see if the Income Generation Model gives a significantly different explanation relative to the variables (Gender and Age) used in the transition matrix. We find that it does. This is unsurprising as most of the extra variables are individually significant. Comparing this with the simulations, we find that the IGM explains more, but in reality it has a limited impact on the simulation properties of the model.

In the second model, we run a Chow-test testing whether there is a difference between the models in two time periods, whether there is a structural change over time. To do this we interact the year equals 2009 dummy with all the variables and test the significance of these interactions. We find that the factors, collectively, are statistically significant indicating that the determinants of employment have changed between the two periods. In a nowcasting exercise, we only have historical data in doing the simulations. The model will therefore not capture changes to the underlying relationship. The question is whether nowcasting improves the model relative to the null assumption of a uniform change in the employment rate applied in the simpler transition matrix approach.

Most economic shocks are asymmetric, resulting in difference shocks across different sectors or population groups. For simplification purposes, we consider an example of an exogenous change in one important labour market variable, industry. In Table 3, we compare the impact of changing industry, utilising the status quo in the Monte Carlo analysis and incorporating individual heterogeneity in the IGM analysis. In this case we find the IGM analysis that incorporates heterogeneity results in a distribution that is not significantly different from the actual result, while the status quo assumption in the Monte Carlo analysis is significantly different for females. The decline in employment disproportionally of younger male workers also captured changes in the manufacturing and construction sectors.

Table 3. Distribution of Industrial Status

	Male			Female		
Industry	Monte Carlo	IGM	Actual	Monte Carlo	IGM	Actual
Agriculture	8.25	8.88	8.80	0.66	0.51	0.36
Construction	13.74	12.94	12.95	6.33	6.74	6.67
Manufacturing	17.40	15.23	15.16	2.57	2.19	2.18
Commerce	17.55	19.91	20.16	21.65	24.61	24.91
Trans and Comm	6.58	5.56	5.60	3.11	1.43	1.37
Public Admin	15.78	17.19	16.96	41.35	41.2	40.94
Health and Ed	15.56	16.38	16.41	17.48	17.31	17.16
Other	5.13	3.91	3.95	6.85	6.01	6.43
Total	100.0	100.0	100.0	100	100	100
Chi-Squared Test of Difference with Actual	0.014	0.000		0.116	0.000	

Earnings

This paper focuses on nowcasting the impact of macro-economic change on income. In Table A (Annex), we report the earnings regression for males and females as a function of personal human capital characteristics and sectoral characteristics from a right skewed distribution of earnings.

Much of the literature on earnings functions use log-normal distributions (Harmon et al, 2002), given the convenient representation of the rate of return to education by the semi-log functional form and the representation of skewness by the log normal distribution. While this is useful in representing the structure of an income distribution, it can cause problems in simulation (O'Donoghue et al., 2009b). The log-normal structure means that an exponent is taken when converting simulated log earnings resulting from this model into earnings. This exponential transformation can have the impact of small positive changes in explanatory variables resulting in large changes in average incomes.

In the case of the nowcasted employment rate above, the (false) increase in the education profile of the work force as a result of a disproportionally higher share of upper secondary educated workers remaining in employment in a downturn, can result in an upward shift in average earnings. This has the risk in a log-normal distribution of increasing the average disproportionally.

Singh and Maddala (2008) developed a model based upon the generalised Pareto and Weibull distribution that provide an estimate of the skewed distribution, which connects individual characteristics to the entire conditional wage distribution and not only the conditional mean like in the standard ordinary least squares (see the Appendix and Biewen and Jenkins (2005) for more details). We report in Table B (Annex) the coefficients of the Singh-Maddala distribution.

In Table 4 we report the ratio of the simulated distribution of earnings using a log normal distribution (OLS) and the Singh-Maddala (SM) distribution for the Monte Carlo (transition probability simulation) and the Income Generation Model (IGM) distribution, with the effect the latter distribution having an improved heterogeneity vis-à-vis industry. We note that the average earnings is higher than the actual in each case. Both of the log-normal models simulate earnings growth at a rate that is statistically significantly higher than the actual for both men and women. The Singh-Maddala distributional regression in the IGM produces simulated values that are not significantly different from the actual, which is desired when simulating.

On the basis that the log-normal distribution may be the source of differences between nowcasting with a transition probability approach or an income generation modelling approach, we also compare the simulation of earnings using the Singh-Maddala model and find that although the male average is not significantly different, the female distribution is. The selection may be the source of this discrepancy.

Table 4. Ratios of Mean Earnings: Approach with OLS vs Approach with SM

	Male	Female
Monte Carlo	1.06	1.19
IGM (OLS)	1.07	1.19
IGM (Distributional Regression)	1.02	1.02
Monte Carlo (Distributional Regression)	1.02	0.93
Statistically Significant Difference in Means		
Monte Carlo	1	1
IGM (OLS)	1	1
IGM (Distributional Regression)	0	0
Monte Carlo (Distributional Regression) (No Industry)	0	1

Turning now to testing the differences in distributional values, we compare the structure of the simulated distribution of earnings for males and females relative to the actual distribution in Table 5. As above in the means comparison, we compare 4 distributions, the distribution of earnings with the transition probability or Monte Carlo scenario, the version with the log-normal regression model and the version with the Singh-Maddala distributional regression model. For the latter we consider both the situation where the explanatory factors are also simulated in the Income Generation Model, or where they remain unchanged, as in the case of the Monte Carlo analysis. The kernel density functions in the appendix highlight relatively small changes to the distribution of earnings in the two periods across different simulation methods.

To compare different distributions we use the Kolgomorov-Smirnov Test, which evaluates the distance between the empirical (simulated) distribution of earnings and the reference (actual) distribution (Table 5). We reject the Null Hypothesis that the distributions are the same, in both the Income Generation Model and the Monte Carlo analysis for both males and females at 95% confidence level, indicating that the simulated distributions of earnings are different. For the Singh-Maddala case, we find that this hypothesis is not rejected for males, giving evidence that the Singh-Maddala method simulation is a better predictor of the actual distribution. This is also found within this method if applied to the Monte Carlo Data.

When we compare the actual distributions for both 2008 and 2009, we find that the Null Hypothesis is rejected at the 90% level for males and the 99% level for females. Therefore the underlying distributions actually changed between the two years in the crisis. The Singh-Maddala method captured the change for males, but not for females. We find that the Singh-Maddala method reproduces the underlying distribution well for the original year. Therefore we conclude that for males, changes in observed characteristics drove the distributional change, while for females, changes in unobserved characteristics drove the change in the distribution.

In model-based nowcasting, we extrapolate into the future on the basis of observed relationships between characteristics and exogenous control totals. Therefore we can have more confidence in a model where the changes are due to observed underlying changes than for changes in unobserved changes. Therefore we have more confidence in the male model than the female model. Improving the latter will require us to understand the drivers of the error component structure in greater detail as in the case of Sologon and O'Donoghue (2014). In summary, the Income Generation Model approach out-performs the Transition Probability or Monte Carlo Method. While it performs similarly in terms of heterogeneity, the Income Generation Model with the Singh-Maddala method performs better for means.

Table 5. Kolgomorov-Smirnov Test of Distributions (p. value of difference)

	Male	Female
Relative to Actual 2009		
Monte Carlo	0.006	0.000
IGM (OLS)	0.006	0.000
IGM (Distributional Regression)	0.133	0.001
Monte Carlo (Distributional Regression)	0.185	0.001
Relative to Actual 2008		
Actual 2009	0.067	0.000
IGM (Distributional Regression)	0.999	0.180

Nowcasted Results

Table 6 reports the confidence intervals around three different income measures, reporting the upper bound and the lower bound for actual incomes collected in the Survey of Income and Living Conditions (EU-SILC). In each case the central estimate falls outside the confidence interval for the previous year. In 2008, the Gini's for gross and disposable income fall below the lower bound of the previous year, while for market income it is borderline not significant. For 2009, the Gini is lower than the lower bound for disposable income and gross income, but higher than the upper bound for market income, reflecting a widening market income and increasing redistribution in the tax-benefit system. As the Gini coefficient typically moves relatively slow, such large inter-annual changes are a symptom of the large changes that occurred during the financial crisis. They are thus a good test of the performance of the simulation model.

Table 6. 95% Confidence Intervals Gini by Income Source

	Lower Bound	Upper Bound	Central
Disposable			
2007	0.289	0.304	0.289
2008	0.282	0.294	0.279
2009	0.263	0.271	0.268
Gross			
2007	0.344	0.365	0.342
2008	0.335	0.354	0.330
2009	0.315	0.330	0.323
Market			
2007	0.471	0.493	0.475
2008	0.464	0.484	0.471
2009	0.479	0.494	0.486

Note: LB – Lower Bound; UB – Upper Bound

Table 7 reports the nowcasted market, gross and disposable incomes for each of the three years. In the case of a simulation one year ahead, we find that in each case the nowcasted Gini coefficient falls within the confidence interval of the actual data. However, when we simulate two years ahead, the nowcasted Gini is significantly different from the actual distribution. For market income, the gap is 1.7 points. It is the same for gross income and smaller at 1.3 for disposable income. This is in part due to differences in the underlying sample that may result from either changes in the underlying population or sample variation.

Table 7. Nowcasted Gini (* significantly different at 95% from baseline)

Disposable	Lower Bound	Upper Bound	Central
2007	0.289		
2008	0.285	0.279	
2009	0.281*	0.268	0.268
Gross			
2007	0.342		
2008	0.337	0.33	
2009	0.340*	0.327	0.327
Market			
2007	0.475		
2008	0.477	0.471	
2009	0.503*	0.493	0.486

While the two-year nowcast is significantly different to the actual value, a more appropriate use of the model from a policy perspective is firstly whether the model captures the direction of the changes, and secondly whether this change in direction has the same level of significance as the underlying trend. In Table 8, we report the significance of changes relative to a 1-year lag or a two-year lag from Tables 6 and 7. In each case we find that the nowcasted Gini for each of the income measures captures the direction of change well. In terms of statistical significance, for disposable income and for gross income, the underlying data was statistically different to a one-year lag and a two-year lag. The trends for nowcasted disposable and gross income follow the same sign and significance. For a one-year lag, for market income, the change is not significant for the actual data. This is also found for the nowcasted market income. However for the two-year lag, the change in 2009, although increasing, is not significantly different from what was observed in 2007. For the nowcasted measure, although capturing the trend, the nowcasted market income distribution exaggerates the actual trend with the simulated value being significantly different to the base year.

Table 8. Statistical significance of change in incomes

	Actual		Nowcasted	
	1-year lag	2-year lag	1-year lag	2-year lag
Disposable				
2007				
2008	1		1	
2009	1	1	1	1
Gross				
2007				
2008	1		1	
2009	1	1	1	1
Market				
2007				
2008	0		0	
2009	1	0	1	1

Table 9 reports the actual and nowcasted redistribution for taxes and benefits, which are almost exactly the same as the redistribution using the actual data. This is due to the fact that tax-benefit systems are largely deterministic, with minor changes due to differences in the underlying market income distribution.

Table 9. Nowcasted Redistribution

	Taxes			Benefits		
Year	2007	2008	2009	2007	2008	2009
2007	0.053			0.133		
2008	0.052	0.051		0.140	0.141	
2009	0.059	0.059	0.059	0.163	0.166	0.160

4. Conclusions

The last two major economic shocks – the Financial and the COVID-19 crises – have changed the landscape of policy modelling by altering the policy-modelling-data nexus. The urgency of the two crises, especially the COVID-19 pandemic, on the one hand, revealed the inadequacy of traditional statistical datasets and models to provide a timely support to the decision-making process made in times of volatility. On the other hand, it accelerated the efforts to obtain timelier, richer and more granular data for policy analysis, challenging the systems to increase transparency, volume, velocity, variety and veracity in data delivery. The speed of the response in data delivery in order to influence the new policies during the COVID-19 crisis signal the introduction of new, more robust and better-organized systems around the globe. The big data revolution and the new programs and systems able to deliver near real time summary data mark a skipped generation in terms of system development.

The data expansion, its transparency, variety and fast availability in the public domain transformed evidence building during the pandemic, enabling mythological advancements. One such advancement occurred in microsimulation modelling, one of the traditional pillars of data-driven policy analysis used to understand the economic consequences of public policies. Microsimulation models usually rely on survey data, which falls short on volume, velocity and veracity given the 2-3 year time lag between collection and availability for analysis, making this data obsolete in times of volatility. The timely availability of big data enabled the publication of real-time calibration totals, which allowed policy modellers to "update" the latest available surveys via nowcasting methods and assess in real time the distributional impact of the crisis.

A variety of different nowcasting approaches have been developed, relying on transition matrix based methods to simulate exits from employment, to calibrate the changes observed in administrative data, but not yet available in household survey data. These methods can be used in a downturn, where there are exits from employment, but face challenges when capturing changes during the subsequent upturn. We propose a method that goes beyond the state-of-the-art by implementing an aligned or calibrated microsimulation approach to generate a counterfactual income distribution as a function of more timely external data than the underlying income survey. It draws upon the inter-temporal equation based dynamic microsimulation modelling literature, allowing for greater heterogeneity and with applicability at different stages of the business cycle.

We evaluate the simulation performance between our approach and the transition matrix approach by undertaking a nowcast for a historical crisis (Great Recession), evaluating each method against an actual change and against each other. We find that nowcasting is a useful methodology that may be used to examine the most up-to-date statistics on labour force participation, income distribution, prices, and income inequality.

With respect to the evaluation of the two methods, we did not find a significant difference between using the Transition Matrix and IGM for calibrating the labour market with respect to the in-work status. We did however found significant differences when the calibration involved structural heterogeneous changes, which affect in turn the simulation of the income sources. When simulating earnings, we detected problems in using the log-normal distribution which can be circumvented by using the more flexible Singh-Maddala distribution. Given the central role that earnings play in the composition of household disposable income, this finding is of high relevance for policy modellers.

In terms of validating our approach, we find that the model replicates the changes in income distribution over one year. Over the longer term, the model is able to capture the trend, but the precision of the levels weakens the further we get from the estimation year.

It would be interesting to extend this analysis to consider the longer term nature of changes, to compare and contrast the gainers and losers over the entire business cycle rather than for a period around the peak. It would also be interesting to decompose these trends by gender and age group, as the force underpinning the economic downturn are differentially quite heterogeneous.

References

Addabbo, T., García-Fernández, R. M., Llorca-Rodríguez, C. M., and Maccagnan, A. (2016). A microsimulation model to measure the impact of the economic crisis on household income. *International Journal of Manpower* Vol. 37 No. 3, pp. 474-493.

Almeida, V., Barrios, S., Christl, M., De Poli, S., Tumino, A., and van der Wielen, W. (2021). The impact of covid-19 on households income in the EU. *The Journal of Economic Inequality*, *19*(3), 413-431.

Álvarez, E., García-Fernández, R. M., and Muñoz, J. F. (2014). A New Method to Estimate the Low Income Proportion: Monte Carlo Simulations. *International Journal of Economics and Management Engineering*, *8*(10), 3281-3284.

Baekgaard, H. (2002). Micro-macro linkage and the alignment of transition processes: some issues, techniques and examples. Technical paper (University of Canberra. National Centre for Social and Economic Modelling); no. 25.

Bargain, O. and Callan, T. (2010). Analysing the effects of tax-benefit reforms on income distribution: A Decomposition Approach, Journal of Economic Inequality 8 (1): 1—21.

Bargain, O. (2012). Decomposition analysis of distributive policies using behavioural simulations, International Tax and Public Finance, 19(5), 708-731 (CNRS: 2, AERES: A)

Beirne, K., Doorley, K., Regan, M., Roantree, B., and Tuda, D. (2020). The potential costs and distributional effect of covid-19 related unemployment in ireland. budget perspectives 2021. paper 1. Dublin: The Economic and Social Research Institute.

Biewen, M. and Jenkins, S. P. (2005). A framework for the decomposition of poverty differences with an application to poverty differences between countries. Empirical Economics, 30(2):331– 358.

Bourguignon, F., Fournier, M, Gurgand, M. (2001). Fast Development with a Stable Income Distribution: Taiwan, 1979-94. *Review of Income and Wealth*, Vol. 47 No. 2, pp 139-163.

Bourguignon, F., Ferreira, F. H., Leite, P. G. (2008).Beyond Oaxaca-Blinder: Accounting for differences in household income distributions. J. Econ. Inequal. 6(2), 117–148.

Brewer, M., Browne, J., Hood, A., Joyce, R., and Sibieta, L. (2013). The short-and medium-term impacts of the recession on the UK income distribution. *Fiscal Studies*, *34*(2), 179-201.

Brewer, M., and Gardiner, L. (2020). The initial impact of COVID-19 and policy responses on household incomes. *Oxford Review of Economic Policy*, *36*(Supplement_1), S187-S199.

Brewer, M., and Tasseva, I. (2020). Did the UK policy response to Covid-19 protect household incomes?. *Available at SSRN 3628464*.

Bronka, P., Collado, D., and Richiardi, M. (2020). The COVID-19 Crisis Response Helps the Poor: the Distributional and Budgetary Consequences of the UK lock-down. EUROMOD Technical Report.

Bruckmeier, K., Wiemers, J. (2012). A new targeting: a new take-up?. *Empir Econ* 43, 565–580. https://doi.org/10.1007/s00181-011-0505-9

Caldwell, S. (1996). 'Health, Wealth, Pensions and Lifepaths: The CORSIM Dynamic Microsimulation Model', In Harding, A. (eds.), 'Microsimulation and Public Policy', North-Holland, Netherlands.

Caldwell S. and R. Morrison, 2000. "Validation of longitudinal microsimulation models: experience with CORSIM and DYNACAN", in Mitton et al. (eds.) Microsimulation in the New Millennium, Cambridge: Cambridge University Press.

Callan T., Keeney, M. J., Nolan, B., Walsh, J. R., (2001)Reforming Tax and Welfare ESRI Policy Research Series 42. Dublin: Economic and Social Research Institute.

Callan, T. and B. Nolan (2010) Inequality and the Crisis: The Distributional Impact of Tax Increases and Welfare and Public Sector Pay Cuts. *The Economic and Social Review*, Vol. 41, No. 4, pp. 461–471

Callan T., C. O'Donoghue and C. O'Neill, (1994). Analysis of Basic Income Schemes for Ireland, Dublin: ESRI Policy Research Series Paper No. 21.

Callan T., van Soest, A., Walsh, J. (2009). "Tax Structure and Female Labour Supply: Evidence from Ireland" Labour: Review of Labour Economics and Industrial Relations, Vol 23 No 1, pp.1-35.

Carta, F. (2020). Timely Indicators for Inequality and Poverty Using the Italian Labour Force Survey. *Social Indicators Research*, *149*(1), 41-65.

Černiauskas, N., Sologon, D., O'Donoghue, C., and Tarasonis, L. (2022). Income inequality and redistribution in Lithuania: the role of policy, labour market, income and demographics. *Review of Income and Wealth*, *68*(S1), S131-S166.

CSO, 2006. EU Survey on Income and Living Conditions. Dublin: the Stationery Office.

Decoster, A., De Swerdt, K., Loughrey, J., O'Donoghue, C., and Verwerft, D. 2008. Matching indirect tax rates on budget surveys for five selected countries. AIM-AP-deliverable. Retrieved March 12, 2009, from http://www.iser.essex.ac.uk/files/msu/emod/aim-ap/deliverables/AIM-AP3.3.pdf.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica: Journal of the Econometric Society*, 1001-1044.

Engler, A. (2020). What all policy analysts need to know about data science. Brookings Report. https://www.brookings.edu/research/what-all-policy-analysts-need-to-know-about-data-science/

Figari, F. and P. Alari and Sutherland, H. (2015). Microsimulation and Policy Analysis. Handbook of Income Distribution. 2. 2141-2221. 10.1016/B978-0-444-59429-7.00025-X.

Figari, F., and Fiorio, C. V. (2020). Welfare resilience in the immediate aftermath of the covid-19 outbreak in italy. *EUROMOD at the Institute for Social and Economic Research, Tech. Rep.*
Fiorio, C. V. and D'Amuri, F. (2005). Workers' Tax Evasion in Italy. Econpubblica Working Paper No. 104, Available at SSRN: https://ssrn.com/abstract=702922 or http://dx.doi.org/10.2139/ssrn.702922

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics, 55(4), 665–676. https://doi.org/10.1016/j.jmoneco.2008.05.010

Hancock, R., Pudney, S., Barker, G., Hernandez, M. and Sutherland, H., (2004). "The Take-Up of Multiple Means-Tested Benefits by British Pensioners: Evidence from the Family Resources Survey," Fiscal Studies, vol. 25(3), pages 279-303

Hardiman, N., Murphy, P., Burke, O. (2009) Compensating Whom For What? Reconsidering the Composition of Public Spending. UCD Geary Institute Working Papers, UCD Geary Institute.

Harmon, C., Oosterbeek, H., and Walker, I. (2003). The returns to education: Microeconomics. *Journal of economic surveys*, *17*(2), 115-156.

Immervoll H., C. O'Donoghue and H. Sutherland, (1999). "An Introduction to EUROMOD". *EUROMOD Working Paper* no. 0/99.

Immervoll, H., Lindström, K., Mustonen, E., Riihelä, M., and Viitamäki, H. (2005). Static data "ageing" techniques. Accounting for population changes in tax-benefit microsimulation. EUROMOD Working Paper, EM7/05.

Jarmin, R.S. and O'Hara, A.B. (2016), BIG DATA AND THE TRANSFORMATION OF PUBLIC POLICY ANALYSIS. *Journal of Policy Analysis and Management*, 35: 715-721. https://doi.org/10.1002/pam.21925

Klevmarken, N. A., 1997. "Modelling Behavioural Response in EUROMOD," Cambridge Working Papers in Economics 9720, Faculty of Economics, University of Cambridge.

Kump, N., and Navicke, J. (2014). *Re-weighting EUROMOD for demographic change: An application on Slovenian and Lithuanian data* (No. EM13/14). EUROMOD Working Paper.

Leventi, C., Navicke, J., Rastrigina, O., and Sutherland, H. (2014). *Nowcasting risk of poverty and income distribution in the EU in 2013* (No. EM11/14). EUROMOD Working Paper.

Levy, L. and Mercador-Prats, M. (2002). "Simplifying the personal income tax system: lessons from the 1998 Spanish reform," *Fiscal Studies*, vol. 23(3), pages 419-443,

Li, J., and O'Donoghue, C. (2013). A survey of dynamic microsimulation models: uses, model structure and methodology. *International Journal of microsimulation*, *6*(2), 3-55.

Li, J., O'Donoghue, C., Loughrey, J., and Harding, A. (2014a). Static models. In *Handbook of microsimulation modelling*. Emerald Group Publishing Limited.

Li, J., and O'Donoghue, C. (2014b). Evaluating binary alignment methods in microsimulation models. *Journal of Artificial Societies and Social Simulation*, *17*(1), 15.

Li, J., La, H. A., & Sologon, D. (2021). Policy, Demography, and Market Income Volatility: What Shaped Income Distribution and Inequality in Australia between 2002 and 2016? *Review of Income and Wealth*, *67*(1), 196-221.

Li, J., Vidyattama, Y., La, H. A., Miranti, R., and Sologon, D. M. (2022). Estimating the Impact of Covid-19 and Policy Responses on Australian Income Distribution using Incomplete Data. *Social Indicators Research*, *162*(1).

Lustig, N., V. M. Pabon, G. Neidhöfer and M. Tommasi, (2021) Short and long-run distributional impacts of COVID-19 in Latin America, CEQ Working Paper 96, Commitment to Equity Institute, Tulane University, June. REVISED.

March, J. S. M., and Marcus, J. S. (2021). Big data versus COVID-19: opportunities and privacy challenges. *European Economic Blogs*. https://european.economicblogs.org/bruegel/2020/scott-marcus-big-data-covid-19-opportunities-privacy-challenges

Matsaganis, M. and Flevotomou, M., (2010). "Distributional Implications of Tax Evasion in Greece," GreeSE - Hellenic Observatory Papers on Greece and Southeast Europe, Hellenic Observatory, LSE, Paper No. 31.

McCarthy, Y and K. McQuinn, (2011). How Are Irish Households Coping with their Mortgage Repayments? Information from the Survey on Income and Living Conditions. *The Economic and Social Review*, Vol. 42, No. 1, pp. 71–94

Morrison, R. (2006). 'Make it so: Event Alignment in Dynamic Microsimulation', DYNACAN Team, Ottawa, Canada.

Navicke, J., Rastrigina, O., and Sutherland, H. (2014). Nowcasting indicators of poverty risk in the European Union: a microsimulation approach. *Social Indicators Research*, *119*(1), 101-119.

Nymand-Andersen, P. (2015), "Big data: the hunt for timely insights and decision certainty: Central banking reflections on the use of big data for policy purposes", IFC Working Paper No 14.

Nolan and Maitre 2000 'A Comparative Perspective on Trends in Income Inequality in Ireland', The Economic and Social Review, Vol. 31, No. 4, October 2000, pp. 329-350.

Nolan, B. 2004, Long-Term Trends in Top Incomes in Ireland, mimeo., ESRI: Dublin.

Nolan, B. and Timothy M. Smeeding, 2005. "Ireland's income distribution ina a Comparative Perspective" Review of Income and Wealth, Blackwell Publishing, vol. 51(4), pages 537-560, December

Nolan, B., B. Maitre, S. Voitchovsky, and C.T. Whelan. (2012). Inequality and Poverty in Boom and Bust: Ireland as a Case Study. *Growing Inequalities' Impacts* Gini Discussion Paper 70.

O'Donoghue C. 2001, "Dynamic Microsimulation: A Survey", Brazilian Electronic Journal of Economics.

O'Donoghue C., 2002, "Redistribution over the Lifetime in the Irish Tax-Benefit System: An Application of a Prototype Dynamic Microsimulation Model for Ireland", Economic and Social Review, Vol. 32, No. 3.

O'Donoghue, C. 2004, "Redistributive Forces in the Irish Tax-Benefit System", Journal of the Statistical and Social Inquiry Society of Ireland.

O'Donoghue, C., Sologon, D. M., Kyzyma, I., and McHale, J. (2020). Modelling the distributional impact of the COVID-19 crisis. *Fiscal Studies*, *41*(2), 321-336.

O'Donoghue, C., Lennon, J., and Hynes, S. (2009a). The Life-cycle Income Analysis Model (LIAM): a study of a flexible dynamic microsimulation modelling computing framework. *International Journal of Microsimulation*, *2*(1), 16-31.

O'Donoghue, C., R. Leach and S. Hynes, (2009b) "Simulating Earnings in Dynamic Microsimulation Models" in A. Harding, P Williamson and A Zaidi (eds.) *New Frontiers in Microsimulation Modelling*. Farnham: Ashgate.

O'Donoghue, C., Loughrey, J., and Morrissey, K. (2013). Using the EU-SILC to model the impact of the economic crisis on inequality. *IZA Journal of European Labor Studies*, *2*(1), 1-26.

O'Donoghue, C. (2014). *Handbook of microsimulation modelling*. Emerald Group Publishing.

O'Donoghue, C., and Loughrey, J. (2014). Nowcasting in microsimulation models: A methodological survey. *Journal of Artificial Societies and Social Simulation*, *17*(4), 12.

O'Donoghue, C., Sologon, D. M., and Kyzyma, I. (2022). *Novel welfare state responses in times of crises: COVID-19 Crisis vs. the Great Recession*. *Socio-Economic Review*.

Pudney S.E., 1992. "Dynamic Simulation of Pensioner's incomes: methodological issues and a model design for Great Britain.", Dept. of Applied Economics Discussion paper no MSPMU 9201, University of Cambridge.

Pudney, S. Hancock, R. and Sutherland, H., (2006). "Simulating the Reform of Means-tested Benefits with Endogenous Take-up and Claim Costs," Oxford Bulletin of Economics and Statistics, vol. 68(2), pages 135-166, 04.

Savage, M., Callan, T., Nolan, B., and Colgan, B. (2019). The Great Recession, austerity and inequality: lessons from Ireland. *Review of Income and Wealth*, *65*(2), 312-336.

Singh, S. K., and Maddala, G. S. (2008). A function for size distribution of incomes. In *Modeling income distributions and Lorenz curves* (pp. 27-35). Springer, New York, NY.

Sologon, D. M., and O'Donoghue, C. (2014). Shaping earnings insecurity: Labour market policy and institutional factors. *Review of Income and Wealth*, *60*, S205-S232.

Sologon, D., Van Kerm, P., Li, J., & O'Donoghue, C. (2021). Accounting for Differences in Income Inequality across Countries: Tax-benefit Policy, Labour Market Structure, Returns and Demographics. *Journal of Economic Inequality*, *19*(1), 13–43.

Sologon, D., O'Donoghue, C., Kyzyma, I., Li, J., Linden, J., & Wagener, R. (2022). The COVID-19 Resilience of a Continental Welfare Regime - Nowcasting the Distributional Impact of the Crisis. *Journal of Economic Inequality.*

Tissot, B. (2017), "Big data and central banking", IFC Bulletin No 44.

Van Soest, A. (1995): Structural models of family labor supply: a discrete choice approach, Journal of Human Resources, 30(1), 63-88.

Whelan, K. (2010) Policy Lessons from Ireland's Latest Depression. *The Economic and Social Review*, Vol. 41, No. 2, pp. 225–254

Appendix - Methodology

(1) Theoretical Framework

The challenge of updating, nowcasting or projecting is to adjust the structure of a dataset to account for changes to the components of the income distribution that change over time from the time when the dataset was collected. In order to aid understanding of this process we introduce a short theoretical framework describing the different processes.

Our objective is to understand the impact of close to real time labour market, policy and price changes on the distribution of welfare. For the purposes of this document, we ignore savings and the inter-temporal and intra-personal distribution of welfare, thus assuming that disposable income and expenditure are equivalent. Similarly, we also ignore welfare derived from non-cash based sources.

We make the assumption that welfare depends the amount of income available or disposable. Disposable income $Y_{D,t}$ at time t depends on market income $Y_{M,t}$, benefits $B(Y_{M,t}, Z_t, \theta_t^B)$ and taxation $T(Y_{M,t}, Z_t, \theta_t^T)$, which in turn depend on personal skills, family characteristics, Z and tax-benefit parameters θ:

$$Y_{D,t} = Y_{M,t} - T\left(Y_{M,t} Z_t, \theta_t^T\right) + B\left(Y_{M,t}, Z_t, \theta_t^B\right)$$

Market income is determined by the receipt of income source i $I_{i,t}$ and amount $Y_{i,t}$, which are both influenced by observable factors (Z), unobservable factors (ε), parameters (θ), and the unit of measurement (t), such as a time period.

$$Y_{M,t} = \sum_{i=1...m} Y_{i,t}^* = \sum_{i=1...m} \left\{ Y_{i,t}\left(Z_{i,t}^Y, \theta_{i,t}^Y, \varepsilon_{i,t}^Y\right) \times I_{i,t}\left(Z_{i,t}^I, \theta_{i,t}^I, \varepsilon_{i,t}^I\right) \right\}$$

As we move over time, these components can change:

- the exogenous characteristics $Z_{i,t}^Y$ or $Z_{i,t}^I$ may change, for example due to aging or changes in skill levels;
- the relationship between these characteristics and either potential income $\theta_{i,t}^Y$ or the receipt of this income $\theta_{i,t}^I$ may change;
- the distribution of the unexplained components $\varepsilon_{i,t}^Y$, $\varepsilon_{i,t}^I$ may change as a result of structural changes in the economy[4].

A potential caveat in using a parametric process is that parametric form of the distributional functions may not fully reflect changes to the actual distribution. However, in the absence of close to real-time non-parametric correction factors, it is not clear what alternative exists.

[4] Although we do not consider it here, the receipt of an income source analogous to the labour participation decision or the labour supply decision is itself endogenous to both the potential wage rate and indeed to the structure of the tax-benefit system. There is a substantial literature analysing the behavioural response to these changes, following the approach by (Bargain & Callan, 2010). However, given the relatively limited short-term behavioural response, the difficulty of separating the impact of a labour demand shock and a labour supply change and as the focus in this document is on the year on year changes, we do not consider the endogeneity of labour participation (etc.). (Bargain, 2012) looked at the impact of a behavioural change within his decomposition considering a growing labour market over a 3 year period.

Thus the distribution of welfare is a function of each source of potential market income $Y_{i,t}$, receipt of the source of market income $I_{i,t}$, taxation and benefits TB_t, prices p_t and demographic characteristics Z_i:

$$D\left(W\left(Y_{i,t}, I_{i,t}, TB_t, p_t \mid i = 1 \ldots m\right)\right)$$

Changes in the Distribution of Welfare

The objective is to model the impact of changes in population characteristics, the labour market and policy on the distribution of welfare. We define the change in distribution of welfare between time *t* and time *t+1* as:

$$\begin{aligned}\Delta_t &= D_{t+1} - D_t = \\ &= D_{t+1}\big(W(Y_{t+1}, I_{t+1}, TB_{t+1}, p_{t+1}, Z_{t+1} | i = 1 \ldots m)\big) - D_t\big(W(Y_t, I_t, TB_t, p_t, Z_t | i = 1 \ldots m)\big)\end{aligned}$$

The changes in the income distribution over time can be modelled in a number of ways. These include:

- Income indexation – the change in the level of income resulting from changes in average wages $Y_{i,t}$;
- Tax-Benefit parametric (θ_t^T, θ_t^B) and structural (T, B) changes;
- Adjustments for changes in the population composition with respect to the demographic structure of the population $Z_{i,t}^j$ and with respect to the prevalence of income sources $I_{i,t}$.

Income Indexation

We define the level of an income source *i* at time *t* as:

$$Y_{i,t}\left(Z_{i,t}^Y, \theta_{i,t}^Y, \varepsilon_{i,t}^Y\right) = \exp\left(\theta_{0,i,t}^Y + Z_{i,t}^Y \theta_{i,t}^Y + \varepsilon_{i,t}^Y\right),\ \varepsilon_{i,t}^Y \sim N(0, \sigma_{\varepsilon_{i,t}^Y}^2).$$

The exponential form is utilized as income generally follows a log-normal distribution. The simplest indexation method relies on a flat growth rate, *r*, similar to growth in GNP per capita or average earnings:

$$Y_{i,t+1} = r * Y_{i,t}\big(Z_{i,t}^Y, \theta_{i,t}^Y, \varepsilon_{i,t}^Y\big) = \exp\big(\ln(r) + \theta_{0,i,t}^Y + Z_{i,t}^Y \theta_{i,t}^Y + \varepsilon_{i,t}^Y\big), \forall i$$

This is equivalent to an adjustment of the intercept of the income model. Depending on data availability it may be possible to utilise an income specific growth rate, r_i:

$$Y_{i,t+1} = \exp\left(\ln(r_i) + \theta_{0,i,t}^Y + Z_{i,t}^Y \theta_{i,t}^Y + \varepsilon_{i,t}^Y\right), \forall i.$$

The growth rate of income sources, however, may also be a function of demographic or labour market characteristics $Z_{i,t}^Y$. For example, the growth rate of earnings may vary between industries, especially during economic changes that impact industries differently. The rate of return may also vary among groups with different levels of education. Furthermore, disparities in income based on gender or race may fluctuate over time. Capturing this effect will require a change (whether linear or non-linear) in the slope of the model:

$$Y_{i,t+1} = \exp\left(\ln\left(r_i\left(Z_{i,t}^Y\right)\right) + \theta_{0,i,t}^Y + Z_{i,t}^Y \theta_{i,t}^Y + \varepsilon_{i,t}^Y\right), \forall i.$$

The unexplained variation in income i, $\varepsilon_{i,t}^{Y}$ can also be decomposed into permanent u_{i}^{Y} and temporary variation $v_{i,t}^{Y}$:

$$\varepsilon_{i,t}^{Y} = u_{i}^{Y} + v_{i,t}^{Y},$$

$$u_{i}^{Y} \sim N\left(0, \sigma_{u_{i}^{Y}}^{2}\right),$$

$$v_{i,t}^{Y} \sim N\left(0, \sigma_{v_{i,t}^{Y}}^{2}\right).$$

Typically it is assumed that $\sigma_{v_{i,t}^{Y}}^{2} = \sigma_{v_{i,t+1}^{Y}}^{2}$. If $\sigma_{v_{i,t+1}^{Y}}^{2}$ or $\sigma_{\varepsilon_{i,t+1}^{Y}}^{2}$ were known, there would not be any need to update the data. For the purposes of this paper, we ignore individual dynamics and income mobility and focus on simulating the cross-section distribution of income.

Changes in Tax-Benefit Rules

The distribution of disposable income is also impacted by alterations to the tax-benefit system. To account for these changes, a static tax-benefit microsimulation model will be necessary to estimate the effects of these modifications (Li et al., 2014).

$$TB_{t} = B\left(Y_{M,t}, Z_{t}, \theta_{t}^{B}\right) - T\left(Y_{M,t}, Z_{t}, \theta_{t}^{T}\right)$$

By adjusting the parameters of the tax-benefit system between period t and $t+1$, it is relatively simple to simulate the impact of parametric changes between periods:

$$\left(\theta_{t}^{T}, \theta_{t}^{B}\right) \rightarrow \left(\theta_{t+1}^{T}, \theta_{t+1}^{B}\right).$$

This results in changes in the simulated benefits and taxes:

$$TB_{t+1} = B\left(Y_{M,t+1}, Z_{t+1}, \theta_{t+1}^{B}\right) - T\left(Y_{M,t+1}, Z_{t+1}, \theta_{t+1}^{T}\right)$$

When also structural policy changes occur between periods, such as a change in the number of tax-bands, the introduction or removal of a tax credit or allowance, the change in eligibility requirements, it is necessary to re-programme the model to incorporate the new policy instruments and the inter-dependence with the existent instruments:

$$B_{t}(\ ) \rightarrow B_{t+1}(\ )$$

$$T_{t}(\ ) \rightarrow T_{t+1}(\ )$$

$$TB_{t+1} = B_{t+1}\left(Y_{M,t+1}, Z_{t+1}, \theta_{t+1}^{B}\right) - T_{t+1}\left(Y_{M,t+1}, Z_{t+1}, \theta_{t+1}^{T}\right).$$

It should be noted however that frequently the simulated benefits and taxes TB_{t}^{*} are not the same as the actual benefits and taxes TB_{t}. This may result from a number of factors including:

- Benefit under take-up: benefits, particularly those that are means tested frequently see take-up rates in reality lower than theoretical eligibility rates. This can be due to information issues, stigma or transaction costs associated with claiming the benefit (See Pudney et al., 2006; Matsaganis & Flevotomou, 2010; Bruckmeier & Weimers, 2012).

- Benefit Fraud: benefit fraud or income under-reporting may see people receiving benefits higher than they are strictly entitled to.
- Income tax evasion: tax evasion can result in an over-simulation of income taxation. While it exists in most countries, it is particularly severe in some countries (Fiorio and D'Amuri 2005; Matsaganis & Flevotomou, 2010)

Information problems may result in the mis-simulation of income taxation or benefits. For example, income tax deductions that depend on information typically not contained in household survey data, such as capital investments, cannot be modelled, thereby overestimating the income tax liability. Similarly, benefits that are paid for a period of time regardless of income changes may result in the household seemingly not being entitled to the benefit, while receiving the benefit in the data.

It is commonplace for tax-benefit microsimulation models to assume 100% take-up of benefits and no tax-evasion. As Figari et al. (2012) argue "microsimulation was developed primarily not as a tool for describing income distributions, but for describing how these distributions might change if the tax-benefit regime were to change." Thus it is assumed that the change in the simulated (*) tax-benefit system is the same as the change in the actual tax-benefit system:

$$TB_{t+1}^{*} - TB_{t}^{*} = TB_{t+1} - TB_{t}$$

Thus we assume that the change in actual inequality is the same as the simulated (*) change in inequality:

$$\Delta_t = D_{t+1}(\ ) - D_t(\ ) = \Delta_t^{*} = D_{t+1}^{*}(\ ) - D_t^{*}(\ )$$

Of course this does not imply that the simulated level inequality is the same as the actual level of inequality in either period:

$$D_t(\ ) \neq D_t^{*}(\ )$$

$$D_{t+1}(\ ) \neq D_{t+1}^{*}(\ )$$

Adjustments for changes in the structure of the population

The other dimensions that can influence the distribution of income are changes in the population structure:

- demographic changes, such as changes to the age structure, the proportion of migrants, or to the education profile;
- labour market changes, such as changes in the occupation or industry profiles;
- changes to the income recipient structure, resulting from labour market changes such as employment income, capital market changes or some changes to the receipt of insurance instruments such as pensions receipt.

The impact of the population and labour market structure is influenced by the population characteristics Z_t and survey weights (because of non-response), $wt_{j,t}$, which themselves are a function of the population characteristics Z_t.

Modelling the change in inequality due to population structural change can be done either by:

- static ageing: changing the survey weights and holding the population characteristics constant, or
- dynamic ageing: changing the population characteristics and holding the survey weights constant.

In this paper, we age our data to be consistent with external calibration data to simulate the impact of macro-labour market, income and policy changes on the income distribution.

(2) Approach

Our objective is to understand the impact of labour market and policy changes on the distribution of welfare. We do this by simulating changes on a historic dataset to the present, also known as nowcasting.

We begin this section by discussing the evolution of nowcasting methods, then we review microsimulation models and their use in income analysis. We discuss next the main differences between static and dynamic ageing and highlight some of the difficulties in using static ageing in volatile economic conditions. We then describe the dynamic modelling approach utilising a system of equations or an income generation model taken from the inequality decomposition literature. We overview a calibration or alignment methodology used to adjust simulations to account for external control totals. Lastly, we describe the tax-benefit model used to incorporate tax-benefit changes.

Nowcasting

Building upon the macro-economic literature (Giannone et al., 2008), there is a growing literature on the use of nowcasting for producing more timely estimates of inequality and poverty (O'Donoghue and Loughrey, 2014). There is an extensive literature on directly modelling poverty incidence (Álvarez et al., 2014). In this paper, however, we are interested in understanding the components that affect different parts of the income distribution, and in deriving income distributional impacts such as poverty and inequality. We do this in an attempt to understand the process by which asymmetric shocks impact different households in different way.

Much of the literature relies on combining more timely labour force data, which has limited or non-existent income data, with income survey information. The EUROMOD nowcasted income distributions (Leventi et al., 2014; Navicke et al. 2014) have had the highest visibility. Their approach uses the European Labour Force Survey to draw employment rates by age, gender and education levels in order to simulate proportional changes of industry-specific employment rates combined with the EUROMOD tax-benefit model to explain the policy outcomes. A number of papers have utilised EUROMOD in the COVID crisis to provide up-to-date distributional implications of the market and policy responses to the crisis (Figari & Fiorio, 2020; Brewer & Gardiner, 2020; Beirne et al., 2020; Brewer & Tasseva, 2021). Lusting et al. (2020) utilised a similar approach randomly allocating a change within a demographic group.

In other words for population sub-group i, with employment rate e_i and random number u, an individual is simulated to be in employment if

$$u < e_i$$

Nowcasting, the target employment probability moves from e_i to e_i^* and the employment rate is recalculated:

$$u < e_i^*$$

Addabbo et al., (2016) extended the parametric perspective of EUROMOD by modelling employment transitions using estimated probit equations: from unemployed and inactive to employed (with/out reduced hours), and from employed to inactive and unemployed. They also modelled the participation in the wage guarantee scheme. Employment income for those simulated to enter the labour market are then imputed using a Heckman selection model. In the case of an employment transition, the probability of employment for sub-group i depends in addition on additional demographic characteristics Z, contained in the probit model, $e(Z)_i$

$$u < e(Z)_i$$

Nowcasting they use "imputed probabilities to obtain probability thresholds according to the relative change in employment status within the strata".

$$u < \frac{e_i^*}{e_i} e(Z)_i$$

Carta (2019) took another approach; instead of taking the labour force status from household surveys, she imputes labour incomes into the Labour Force Survey. In doing that she draws upon recent labour distributions and adds modelled income utilising Mincerian wage equations. In other words, the paper does not simulate the nowcasted employment rate e_i^*, but the wage rate. While this improves knowledge about the heterogeneity of employment such as within household correlations and correlations with demographic factors, the paper has a different focus considering only wage income rather than all market incomes and consequential disposable incomes. The method, however, has merit and would be worth exploring the potential of simulating other income sources in a Labour Force Survey. We leave this to further work.

Li et al. (2020) take a semi-parametric perspective, drawing upon the methodology of (DiNardo et al., 1996).[5] They use the monthly Australian Longitudinal Labour Force Survey (LLFS) data, together with administrative payroll data. They utilise changes in the LLFS to reweight the labour market characteristics at the intensive and extensive margins of the labour market in the income survey along the lines of demographics, employment and household characteristics.

While convenient in multi-country simulations, the use of Monte-Carlo simulations based upon cell-specific probabilities may ignore some of the important heterogeneity exhibited in a crisis. For example, family status may be an important driver. Similarly, the situation of partners within a family may be correlated as identified in Carta (2019). Carta (2019) avoids issues associated with intra-household variations or sectoral biases by using recent labour force survey data. However, ignoring the impact of other non-labour force characteristics and, in particular, the impact of public policy as an insulating mechanism is an issue. The question is whether it is harder to simulate labour market changes in a model that contains the full range of incomes and policies or vice versa.

[5] Kump, & Navicke (2014) also use a reweighting approach to incorporated changed demographic change.

Reweighting or semi-parametric approaches are strong in the sense that they avoid distributional assumptions. However, like other static-reweighting procedures they rely on the existence of sufficient sample sizes. Increasing the number of dimensions in the analysis increases the risk of relying on cell weights with small numbers. Utilising the Di Nardo et al. (1996) approach has a higher data requirement than calibrated approaches, requiring access to micro data to generate counter-factual weights. Working at close to real time this may not be feasible. As a result, in this paper, we focus on calibrated approaches.

In the case where reweighting or semi-parametric approaches prove unfeasible, a parametric approach may be a more pragmatic approach. In this paper we consider in more detail a parametric approach akin to the Addabbo et al. (2016) methodology, but in addition drawing upon the alignment aspects of the dynamic microsimulation literature, dynamic ageing. We applied this approach for looking at the impact on the COVID crisis in Ireland (O'Donoghue et al.,, 2020). In this paper, we discuss in detail the methodology and we validate it using a short to medium term horizon.

Dynamic ageing vs. static ageing

Dynamic ageing (as described in Li and O'Donoghue, 2013) involves the estimation of a system of econometric equations which simulate changes in the population. This method is referred to as dynamic because it models changes in the distribution rather than individual income dynamics. The process starts with a sample of households with fixed underlying characteristics, which are then altered through a dynamic simulation mechanism to produce different distributions based on expected future characteristics or present characteristics in the case of nowcasting. This approach involves estimating a system of equations that reflect the distribution of income sources.

Dynamic ageing has both advantages and disadvantages compared to static ageing, so Pudney (1992) suggests that neither approach should be used exclusively. Dynamic ageing, which focuses on individuals, does not account for market-level processes such as labor demand. Additionally, dynamic ageing requires a large amount of data and modeling resources to jointly estimate all necessary processes. Typically, dynamic ageing models at the individual level and deals with inter-dependencies between individuals in a conditional manner rather than jointly.

However, static ageing has some theoretical limitations. Firstly, it cannot be used when there are no individuals in a specific state in the sample. If there are a limited number of instances of a certain household category, a high weight may have to be applied, resulting in unreliable predictions. Over time, shifting demographic and economic trends may require an increasing weight to be placed on population segments with a small number of cases in the sample (Klevmarken 1997). Secondly, static ageing methods are suitable for short to medium-term forecasts where changes in the population structure are minimal. But, over longer periods or during turbulent times, it may be challenging to use static ageing due to the changing characteristics of the population.

Immervoll et al. (2005) identify a number of issues:

- controlling for changes in aggregate group sizes may not improve the match for distributional patterns if much of the changes occur within the group;
- ageing techniques should not be mechanically used for different time periods as the suitability of a specific set of alignments will largely depend on changes in the population structure or the tax-benefit system;

- modifying the statistical weights in a dataset to include information about the target population may lead to a distortion of the original weight information. Specifically, adjustments to marginal distributions often result in distortions of joint distributions. While "minimum distance" methods attempt to retain as much information as possible, the probability of such distortions increases with the number of dimensions used for re-weighting and the magnitude of change in each dimension. In such a situation, ageing-techniques may render the “unadjusted” data non-representative of the target population.

Some of the criticisms of static ageing are applicable to dynamic ageing too. From a user’s point of view however, calibrated dynamic ageing has a twofold advantage compared to static ageing:

- it can handle more easily many different control totals that are required to reflect the changes within an economic crisis;
- the user has more control, as the external totals are simulated in a conditional manner and the selection is a function of a statistical distribution.

Due to the significant and fast changes in the economy's structure, and for the reasons stated above, we opt to use a dynamic ageing approach to adjust our income distribution data to reflect macro-economic changes.

Income Generation Models

The dynamic ageing methodology has been used in both dynamic microsimulation and the related pensions literature (Li and O’Donoghue, 2013), and in the economics literature from an inter-temporal (Černiauskas et al. 2022; Bourguignon et al. 2001) and cross-country perspective (Bourguignon et al. 2008; Sologon et al., 2021). Both involve the specification of an income generation model that describes the income generation process, built around the theoretical model described in section 2.

Modelling binary events, such as being in-work, is done using a logit model due to its computational simplicity. To apply the probabilities estimated from logistic models in a Monte Carlo simulation, we generate a set of random numbers to predict the actual dependent variable in the base year.

We define our logit model of the employment rate e_i as follows:

$$y_i^* = ln(e_i) = ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \sum\nolimits_k \beta . X_i^k + \varepsilon_i$$

such that

$$y = 1 \; if \; y_i^* > 0$$

In order to create the stochastic term ε_i, we use:

$$\varepsilon_i = ln\left(\frac{u_i}{1 - u_i}\right)$$

such that

$$y = 1 \; if \; u_i < ln^{-1}\left(\beta_0 + \sum_k \beta . X_i^k\right) = e_i$$

A value of u_i that satisfies this is

$$u_i = (Y = 1) * (r * e_i) + (Y = 1) * \left(e_i + r * (1 - e_i)\right)$$

Where r is a uniform random number.

Once determined whether an individual is in-work or not, their work status, employee, farmer, etc., multi-category choices like occupation and industry are modelled using a reduced form multinomial logit model. This type of model is used when the explanatory variables are not specific to a particular choice.[6] The disturbance terms for multi-category dependent variables such as occupation or industry are derived from multinomial logit models in several steps. We first generate a set of random variables for counterfactual choices using the extreme value distribution:

$$v_j = -ln\left(-ln(u)\right)$$

where u is a uniform random number and j is choice j, not the actual choice chosen by the individual in the original data. Next, we choose a random variable from the extreme value distribution, v_i for the actual choice *I,* such that:

$$XB + v_i > XB + v_j, \forall j \neq i.$$

After establishing the labour force characteristics of each individual, we model the income variables using ordinary least squares:

$$Y_i = \exp\left(X^* B + \varepsilon\right).$$

The disturbance term is normally distributed, generated directly from the data for those with observed incomes, or generated stochastically for those without a specific income source in the data.

Given the central role of earnings in the composition of household income, the income generation model (IGM) uses the flexible Singh-Maddala parametric distributional regression which connects individual characteristics to the entire conditional wage distribution and not only the conditional mean like in the standard ordinary least squares (see Biewen and Jenkins (2005)). Individual earnings are given by

$$w_{hi} \; = \; F_{X=z}^{-1}(v_{hi}) = b(z)[(1 - v_{hi})^{-\frac{1}{q(z)}} - 1]^{\frac{1}{a(z)}},$$

where v_{hi} is a random term uniformly distributed, $q(z)$ is a shape parameter for the 'upper tail', $a(z)$ is a shape parameter for the 'spread' affecting both tails of the distribution, and $b(z)$

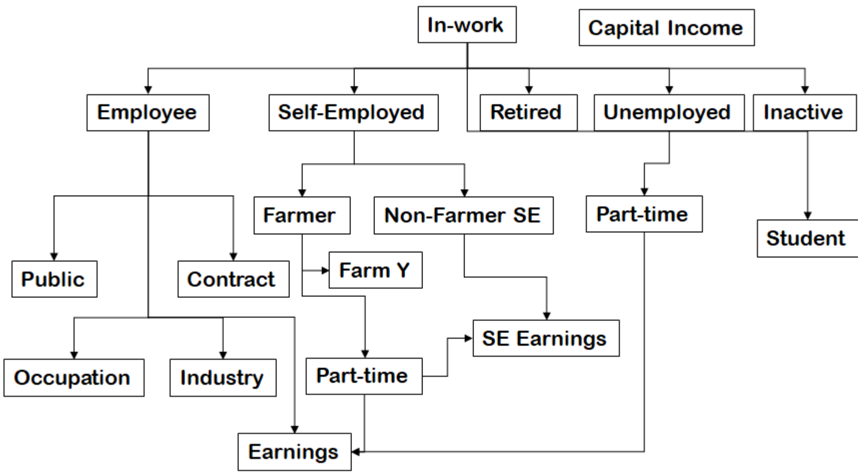
[6] There is extensive research on using choice-specific models to model multi-category choices, such as in the case of structural labour supply equations (as discussed in Van Soest, 1995; and Callan et al., 2010). However, in our approach, we utilize a calibration method described below, which overrides the behaviour captured by these models.

is a scale parameter. The parameters a, b, and q are linearly related to individual characteristics z, which encompass labor market factors (such as occupation, industry, and sector).

Most microsimulation models apply the standard ordinary least squares for modelling earnings and wages. By applying the flexible parametric distributional regression in a microsimulation context, we follow the latest developments in field (Sologon et al. 2021). We bring evidence the distributional regression performs better than the standard OLS.

The diagram in Figure 1 shows the system of equations at the core of the income generation model. We estimate these models and save the parameter estimates and the residuals. These will in turn be used in the simulations, calibrations and alignments.

Figure 2. Income Generation Model



Simulation, Calibration and Alignment

We use a calibration method to simulate changes in exogenous control totals to project the distribution forward over a period using a single set of estimated parameters, a single set of explanatory factors and a single set of residuals.[7]

In this paper, instead of focusing on the tracking of individuals moving through different states, as is common in dynamic microsimulation modeling, our objective is to reproduce cross-sectional distributions of income, similar to Bourguignon et al. (2001). This enables us to use a simpler error component structure without the need for time-based components and base our estimations on cross-sectional data. However, this approach limits us to analyzing the impact of external changes on the distribution shape, in relation to groups such as income quintiles or labor market groups, rather than examining changes in the welfare of specific individuals or analyzing the number of winners and losers.

To accomplish this, we utilize external control totals that are more recent than micro-income survey data and use a calibration or alignment technique, as described in O'Donoghue et al.

[7]Also, unlike the dynamic microsimulation modelling literature we do not model transitions in this analysis as the SILC data only has limited panel information. Therefore, in our objective to simulate changes in the labour market and wage rates as in the case of Bourguignon et al. (2002), we simulate cross-sectional models.

(2008), Caldwell (1996), and Morrison (2006). The purpose of calibrating the microsimulation model is to make sure that the simulated output aligns with exogenous totals (Baekgaard, 2002). In our model, we employ three types of alignment for binary discrete data, discrete data with more than two choices, and continuous data.

Binary choice models are calibrated by ranking y^* defined in (1) above and selecting the highest N cases from our external control totals of for example the numbers in work N_i:

$$y = 1 \; if \; rank(y^*) < N_i$$ [8]

This approach overcomes a number of simulation issues found in other calibration approaches (Li and O'Donoghue, 2014).

In multiple-choice models, a similar method is developed, ranking y_j^* for each choice j in turn to be consistent with externally defined N_j Income variables are adjusted by up-rating using group specific income growth rates defined above.

In relation to simulating income, we adjust the explanatory variables, such as labour force participation variables, through the alignment methodology outline above. However, unlike Bourguignon et al. (2001) who used information from historical surveys on the standard deviation of earnings over different periods, we do not directly adjust the earnings distribution. We however use sector and occupation specific earnings growth indices taken from the National Employment Survey to up-rate earnings.

Tax-Benefit System

Disposable income, defined as income after direct taxation and social benefits is calculated through the use of a static tax-benefit model, which simulates the main direct tax and transfer instruments:

- Income Taxation
- Social Insurance Contributions (Employee, Self-Employed and Employer)
- Income Levies
- Family Benefits
- Social Assistance Benefits
- Social Insurance Benefits.

The Irish tax-benefit system belongs to the Anglo-liberal welfare state category, where social transfers aim primarily at reducing poverty through means-tested benefits or flat-rate insurance benefits.[9] Our tax-benefit model simulates the amount of social insurance benefits paid, assuming eligibility is based on receipt in the data. The benefit system lacks any earnings-related components. The income tax system has a two-rate schedule and allows for optional joint filing, with partial transfer of tax bands and credits. Instead of allowances, tax credits are used in the tax system.

[8] It should be noted that as N_i is a population value and the survey contains a sample of the population, the rank used for calibration is the weighted rank.

[9] For a broad description of the structure of the Irish tax-benefit system, see O'Donoghue (2004).

Appendix - Tables

Table 10. Earnings Regression (OLS)

	Male		Female	
	Beta	SE	Beta	SE
Age	0.1519***	0.0087	0.168***	0.008985
Age Squared	-0.0016***	0.0001	-0.0017***	0.000103
Married	0.1834**	0.0876	0.0258	0.063359
Single	-0.1833*	0.0963	0.209***	0.075403
Years of Education	-0.0327**	0.0134	-0.0647***	0.015089
Years of Education Squared	0.0047***	0.0008	0.0072***	0.000938
Part-Time Work	-0.5484***	0.0629	-0.5817***	0.041256
Occupation 2	-0.0283	0.0600	0.2528***	0.065539
Occupation 3	-0.0112	0.0775	0.1702*	0.089297
Occupation 4	-0.2031***	0.0538	-0.0288	0.050875
Occupation 6	-0.1562	0.1695	-0.6067	0.620079
Occupation 7	-0.0294	0.0557	-0.3484	0.246789
Occupation 8	-0.1763**	0.0707	-0.6242***	0.165941
Construction	0.3535***	0.1040	0.5404*	0.318275
Manufacturing	0.4723***	0.1065	-0.2046	0.326174
Commerce	0.4334***	0.1038	0.1813	0.311317
Trans & Comm	0.431***	0.1160	0.3887	0.329465
Public Admin	0.5565***	0.1051	0.3298	0.310587
Health and Ed	0.5232***	0.1063	0.3513	0.312458
Other	0.1762	0.1256	0.2391	0.317548
Constant	6.2432***	0.2313	5.3511***	0.371502
N	2111		2218	
R2	0.4212		0.3837	

Table 11. Earnings Regression (Singh Maddala Distribution)

a				
Age	0.1789***	0.0108	0.0513***	0.005812
Age Squared	-0.0021***	0.0001	-0.0006***	6.17E-05
Married	-0.6249*	0.3426	0.097**	0.040959
Single	-1.3955***	0.3702	0.0813	0.053355
University Educated	-0.1622**	0.0793	0.1343***	0.030064
Number of Children	-0.018	0.0262	-0.0427***	0.010501
Part-Time	-0.1685***	0.0647	-0.173***	0.031791
Public Sector	0.3274***	0.1069	0.8789***	0.075663
Constant	0.0409	0.4372	0.5343***	0.129476
b				
University Educated	14445.25***	4188.0	13996.44	9246.5
Married	4283.249	3008.2	18776.44**	8335.6
Single	39693.13**	15752.5	36511.4**	17081.8
Number of Children	-856.8455	894.6	-6315.88*	3482.5
Age Squared	-3.5481***	1.3	-10.3774**	4.5
Agriculture	-10570.95***	3733.1	-190009.3**	81466.5
Manufacturing	3828.695**	1930.6	-50159.07	38127.4
Public Admin	11272.34***	3082.9	-260797.9***	90593.9
Health and Ed	6432.939***	2473.4	125773.4***	42872.8
Part-Time	-27289.08***	2990.0	-28563.06**	11544.9
Occupation 1	25318.45***	3459.8	146847.3***	42079.3
Occupation 2	23412.11***	3424.7	92998.89***	24379.0
Occupation 3	14949.79***	3648.6	85585.9***	27909.0
Occupation 4	2060.542	2118.0	18788.13**	7599.3
Occupation 6	2026.021	3402.3	-216904.3**	101213.0
Occupation 7	7505.325***	2095.7	-6412.183	29125.8
Occupation 8	2409.493	2322.3	44593.35	29459.7
Irish Citizen	6859.502***	1995.5	25449.15**	9948.4
Constant	37501.17***	4752.7	352250.9***	111314.7
q				
Age	-0.2432***	0.0651	-3.9896*	2.179867
Age Squared	0.002***	0.0006	0.0242*	0.013903
Married	0.1525	0.2994	10.6949	7.594389
Single	7.9425**	3.1030	13.8442	14.4926
University Educated	-0.4716**	0.2066	-12.3137	8.377003
Number of Children	-0.2132***	0.0530	-3.2865	2.599123
Part-Time	-0.5484**	0.2696	27.806	17.60211
Public Sector	0.4062*	0.2307	-40.1872**	15.83405
Constant	8.8071***	1.8218	208.979**	103.0003
N	2111		2218	