******

***Research Outputs: Population Estimates Compiled from Administrative Data Only (PECADO)***

*4th December 2018*

*John Dunne, CSO*

*Comments and queries to* [*adc@cso.ie*](mailto:adc@cso.ie)

Table of Contents

[Disclaimer 2](#_Toc531008945)

[Main Points 3](#_Toc531008946)

[Contributing Data Sources 5](#_Toc531008947)

[Table 1: List of data sources contributing to SPD by year (Y/N indicator) 5](#_Toc531008948)

[Research Outputs 5](#_Toc531008949)

[Table 2: PECADO annual population estimates by Sex, Statistic and Year 6](#_Toc531008950)

[Table 3: PECADO annual population estimates by Age in April of reference year, Statistic and Year 6](#_Toc531008951)

[Comparison of Research Outputs with Census Counts 7](#_Toc531008952)

[Table 4: Comparison of PECADO annual population estimates and Census (Usually Resident and Present) counts 7](#_Toc531008953)

[Figure 1: Comparison of PECADO Research Outputs and Census counts by age and gender, 2011 8](#_Toc531008954)

[Figure 2: Comparison of PECADO Research Outputs and Census counts by age and gender, 2016 8](#_Toc531008955)

[SPD Coverage Rate 9](#_Toc531008956)

[Table 5: SPD 'Signs of Life' Coverage of the Population, 2011 - 2016 9](#_Toc531008957)

[Next Steps 9](#_Toc531008958)

[Further reading 9](#_Toc531008959)

# Disclaimer

These Research Outputs on Population Estimates Compiled from Administrative Data Only (PECADO) are not official statistics. Rather they are disseminated as outputs from research into a different methodology to that currently used in the production of official population estimates.

It’s important that the information and research presented here be read alongside the outputs to aid interpretation and avoid misunderstanding. These outputs must not be reproduced without this disclaimer and warning note.

# Main Points

This research is part of a larger CSO and NSB strategic goal to be able to compile, at a lower cost to the taxpayer, detailed ‘census like’ population statistics from administrative data on an annual basis, at detailed geographic level. This will be a requirement of the European Statistical System from 2024 onwards. Administrative data is sourced from systems operated by other public authorities (i.e. welfare, tax, education etc.). The first milestone in achieving this objective is to compile reliable and robust population estimates at a state level in a safe and reliable manner. The research outputs presented here are the result of work undertaken to date to achieve this first milestone. This is the PECADO project.

The PECADO project is particularly innovative as no country without a Central Population Register (CPR), as far as we know, has yet to compile population estimates from administrative data sources without undertaking a significant coverage survey in the field as part of the methodology.

The target population concept used for this research “is any person usually resident in the state at any given part of the reference year”. “Usually resident” is defined as residing in the state, *or intending to reside in the state*, for a period of 12 months or more. This population concept includes anybody that resided in the state and then emigrated in the reference year as well as any person who immigrated into the state with the intention of residing here in the reference year. In practice, the PECADO project has used Signs of Life (SoL) on public administration systems as a proxy for usually resident, and as such, persons residing in the State for less than 12 months but engaging with public administration systems will be included in the PECADO estimates. For convenience, the calendar year rather than a point in time is used as the reference period.

The research outputs present population estimates for the years 2011 to 2016 broken down by gender and single year of age (along with precision estimates).

The underlying methodology can be summarised as follows

* a (pseudonymised) statistical population dataset (SPD) is first compiled using a ‘signs of life’ approach that only include persons in the dataset where there is evidence that they have engaged with the underlying public administration systems in the reference year. By definition, the SPD will contain under-coverage, but no over-coverage. In practice, however there is always a small risk of erroneous records being included.
* the counts from the SPD are then adjusted using Dual System Estimation (DSE) methods, also known as capture recapture methods, to obtain population estimates using another administrative data list that has been deliberately excluded from the SPD.
* The administrative data list used to adjust the SPD is compiled from the Driver Licence Database and includes only those persons that apply for or renew their driver licence in the reference year.
* the underlying DSE assumptions underpinning the use of this second list in DSE methodologies are validated by swapping in a survey dataset and comparing the two sets of population estimates
* the SPD is further evaluated for erroneous records (possible overcount) using an extension to DSE methodologies called Trimmed Dual System Estimation, Zhang and Dunne (2018)

A Statistical Population Dataset (SPD) is a dataset that typically contains a record for each member in the population.

Innovative aspects of the PECADO project include

* taking a ‘signs of life approach’ to creating the SPD and, as such, eliminating the need to address over-coverage issues
* using an administrative data source to adjust for undercount in the SPD
* applying an extension of the DSE methodologies (TDSE - Trimmed Dual System Estimation) to look for erroneous records on the SPD.

More information on the underlying methodologies are available in the accompanying draft working paper where the methods and estimates are explored in more detail.

# Contributing Data Sources

The SPD for a given reference year is created from pseudonymised versions of the contributing data sources where a Protected Identifier Key (PIK) has been deployed for each person. The PIK is then used to link the pseudonymised data sources to create a record for each person on the SPD. The SPD contains demographic information (gender, nationality, month of birth) along with summary information from the underlying data source (typically a 0/1 indicator for activity). The PIK protects a person’s identity but also enables linking across data sources and over time. The PIK used in this case is derived from the PPSN and enables high quality deterministic matching thus significantly reducing/eliminating linkage error.

## Table 1: List of data sources contributing to SPD by year (Y/N indicator)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data source \ Year | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|  |  |  |  |  |  |  |
| Social Welfare excluding pensions | Y | Y | Y | Y | Y | Y |
| P35 records (Revenue) | Y | Y | Y | Y | Y | Y |
| Post Primary Pupils | Y | Y | Y | N | N | N |
| Income Tax (Revenue) | Y | Y | Y | Y | Y | N |
| Child Benefit - Child | Y | Y | Y | Y | Y | Y |
| Child Benefit - Parent | Y | Y | Y | Y | Y | Y |
| ECCE Child | Y | N | N | N | N | N |
| Higher Education | Y | Y | Y | Y | Y | N |
| Further Education | Y | Y | Y | N | Y | N |
| Social Welfare Short term benefits | Y | Y | Y | Y | Y | Y |
| Primary Care Reimbursement Service (PCRS) | Y | Y | Y | Y | Y | Y |
|  |  |  |  |  |  |  |

The distribution of each data source can also be explored by age, sex and year from the interactive visualisation tool. This tool does not provide a definitive count of activity on administrative systems. It only provides a count of those records that have contributed to the compilation of the population estimates (at the time of compilation).

See/link Lyndas explorer here

<http://adc/productdevelopment/AdminPopTree/PECADO_All_Nationalities/PecadoExplorer.html>

# Research Outputs

From the PECADO project there were 5,038,640 persons estimated as being usually resident in the State during 2016.

The following tables give breakdowns of the PECADO population estimates by gender and age group for the years 2011 to 2016.

## Table 2: PECADO annual population estimates by Sex, Statistic and Year

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  |  |  |
|  | PECADO estimate (Persons) | | |  |  |  |
|  | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
| Both sexes | 4,811,020 | 4,828,990 | 4,896,230 | 4,925,380 | 4,992,260 | 5,038,640 |
| Male | 2,421,310 | 2,433,930 | 2,466,580 | 2,478,580 | 2,514,440 | 2,539,120 |
| Female | 2,389,710 | 2,395,060 | 2,429,660 | 2,446,800 | 2,477,810 | 2,499,520 |
|  |  |  |  |  |  |  |

## Table 3: PECADO annual population estimates by Age in April of reference year, Statistic and Year

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PECADO estimate (Persons) | | | | | |  | |  | |  | |
|  | 2011 | | 2012 | | 2013 | | 2014 | | 2015 | | 2016 | |
| All ages | | 4,811,020 | | 4,828,990 | | 4,896,230 | | 4,925,380 | | 4,992,260 | | 5,038,640 | |
| 0 - 4 years | | 370,910 | | 372,030 | | 368,620 | | 361,780 | | 354,320 | | 345,970 | |
| 5 - 9 years | | 326,670 | | 329,600 | | 338,910 | | 346,350 | | 353,830 | | 362,380 | |
| 10 - 14 years | | 306,200 | | 309,280 | | 313,440 | | 315,230 | | 318,360 | | 321,440 | |
| 15 - 19 years | | 283,420 | | 283,290 | | 289,670 | | 292,220 | | 297,820 | | 303,570 | |
| 20 - 24 years | | 322,050 | | 308,500 | | 306,140 | | 303,450 | | 301,420 | | 297,010 | |
| 25 - 29 years | | 417,940 | | 389,900 | | 372,060 | | 357,420 | | 349,890 | | 345,140 | |
| 30 - 34 years | | 435,720 | | 436,690 | | 439,080 | | 428,220 | | 421,180 | | 410,440 | |
| 35 - 39 years | | 391,530 | | 392,270 | | 396,650 | | 400,760 | | 412,730 | | 423,160 | |
| 40 - 44 years | | 349,510 | | 356,410 | | 365,790 | | 371,680 | | 381,120 | | 387,910 | |
| 45 - 49 years | | 317,470 | | 321,170 | | 326,030 | | 328,800 | | 335,650 | | 343,890 | |
| 50 - 54 years | | 281,570 | | 286,660 | | 293,890 | | 301,250 | | 310,150 | | 312,880 | |
| 55 - 59 years | | 249,960 | | 253,600 | | 259,290 | | 263,570 | | 270,210 | | 277,840 | |
| 60 - 64 years | | 222,150 | | 225,880 | | 231,680 | | 234,810 | | 239,540 | | 241,640 | |
| 65 - 69 years | | 179,670 | | 190,400 | | 197,980 | | 208,230 | | 214,560 | | 218,080 | |
| 70 - 74 years | | 132,300 | | 137,260 | | 143,570 | | 151,230 | | 159,640 | | 169,030 | |
| 75 - 79 years | | 101,900 | | 105,510 | | 110,760 | | 113,550 | | 118,480 | | 120,640 | |
| 80 - 84 years | | 69,180 | | 72,040 | | 76,940 | | 79,340 | | 82,520 | | 85,130 | |
| 85 years and over | | 52,860 | | 58,500 | | 65,740 | | 67,500 | | 70,860 | | 72,510 | |
|  | |  | |  | |  | |  | |  | |  | |

Disaggregation of the PECADO estimates, cross classified by the classification variables gender and age, is available via Statbank style tables at <http://www.cso.ie>/xxx . The tables also include counts of records used from underlying administrative systems along with precision estimates.

The results can also be visualised using population trees with the PECADO interactive explorer. See

<http://adc/productdevelopment/AdminPopTree/PECADO_All_Nationalities/PecadoExplorer.html>

# Comparison of Research Outputs with Census Counts

As can be expected, there will be differences between the PECADO research outputs presented here and Official CSO population estimates as defined by the Census.

Reasons for differences can be attributed to differences in the underlying population concepts and or possible violations in the assumptions underpinning the PECADO estimates (including the possibility of erroneous records being present in the underlying data sources).

There is a difference of 5.2% in the 2011 overall figure and this rises to 6.3% in the 2016 overall figure for the usually resident population.

## Table 4: Comparison of PECADO annual population estimates and Census (Usually Resident and Present) counts

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 2011 | 2016 | difference | % difference |
|  |  |  |  |  |
| PECADO Estimate | 4,811,020 | 5,038,640 | 227,620 | 4.7 |
| Census Count | 4,574,890 | 4,739,600 | 164,710 | 3.6 |
| difference | 236,130 | 299,040 |  |  |
| % difference | 5.2 | 6.3 |  |  |
|  |  |  |  |  |

The population concept underpinning the PECADO annual population estimates relates to persons resident in the State at any timepoint in the year, while the population concept underpinning the Census Counts presented here relates to persons usually resident in the State and present on Census night. The PECADO project uses SoL as a proxy for being counted as usually resident while Census usually resident concept is based on a 12 month criteria.

The Census of Population counts are the Official CSO figures.

A further exploration of the differences by age and gender is shown in the graphs below for 2011 and 2016. PAR (blue) refers to the counts from the Statistical Population Dataset. DLD (green) refers to the records used from the Driver Licence Database. Match (grey)refers to the records that were matched between them. Population estimates are in black and can be compared against the Census usually resident estimate in Red.

The biggest differences occur with respect to male adults aged between 20 and 50.

## Figure 1: Comparison of PECADO Research Outputs and Census counts by age and gender, 2011



## Figure 2: Comparison of PECADO Research Outputs and Census counts by age and gender, 2016



# SPD Coverage Rate

The SPD coverage rate for the population is estimated at just over 90%. The table below provides information on the coverage rate for each year.

## Table 5: SPD 'Signs of Life' Coverage of the Population, 2011 - 2016

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
| SPD | 4,397,770 | 4,424,370 | 4,533,430 | 4,541,630 | 4,611,800 | 4,473,900 |
| PECADO Estimate | 4,811,020 | 4,828,990 | 4,896,230 | 4,925,380 | 4,992,260 | 5,038,640 |
| % Coverage | 91.4 | 91.6 | 92.6 | 92.2 | 92.4 | 88.8 |

# Next Steps

1. Further work is planned to explore the robustness of the PECADO system and the underlying estimates.
2. Work will commence on developing the methodologies and systems to disaggregate the PECADO estimates by geographical area and household composition.
3. Promote the development of underlying data sources – EIRCODES, metadata

# Further reading

PECADO Project methodology

Zhang and Dunne (2018) Trimmed Dual System Estimation, Chapter 17 in “Capture-Recapture Methods for the Social and Medical Sciences” published by CRC Press.