Obtaining taxable income data from admin data

1. Data Requirements:

In order to produce the taxable income timeseries from admin dataset, we would need the following three datasets:

- **IR3: Individual Income Tax Returns:**
  The IR3 is a self-completed tax return form required for individuals in New Zealand who earn income beyond salary, wages, interest, dividends, or taxable Māori authority distributions.

    - IRD Returns IR3 Key Points (Data up to 2022):
      This dataset highlights active records with non-zero income from partnerships, self-employment, or shareholder salaries. It is designed for mandatory annual income tax filings relevant to specific businesses and individuals.

    - IR3 Full Dataset (Including IR3: 2013–2020 and IR3: 2000–2014):
      Sourced from the IRD's START system (which replaced the earlier FIRST system), this comprehensive dataset was provided directly to the Integrated Data Infrastructure (IDI) without preprocessing. As a result, it includes additional variables not typically retained in "IRD Returns IR3 Key Points". Note that there is a minor overlap in data for the years 2013 and 2014 between the two periods covered (2000–2014 and 2013–2020).

  The above are combined to construct a more comprehensive IR3 dataset.

- **IR Autocalc information:**
  this data shows the information IR uses to calculate Automatic Assessments. This auto-calc process applies to people whose income is only salary, wages, interest or dividends. This income information is received by IR from employers and payers of investment income.

- **IR Personal Tax Summary (PTS):**
  this contains the personal tax summary (PTS) information. It is a pre-2019 process (now largely replaced by Auto-Calc) where the IRD sent people a summary of your income and tax paid, which you could review and confirm or adjust.

- **Income by tax year summary**:

This table contains a detailed summary of income data by tax year. There is one record per individual per year, and each record contains all income sources that a person may receive. This data is derived from the IR EMS dataset.

Note that this table is transformed from the table “Income by tax year”, which is comprised of all records in the Employer Monthly Schedule (EMS), plus additional records from the IR3, IR4S and IR20 tax forms. As the information on the IR3, IR4S and IR20 tax forms relate to the tax year, income sources from these sources that relate to individuals have been incorporated into this table. The records have then been arranged into the granularity of one record per payee/payer relationship, per income source, per income

- **Concordance:** the concordance table gives links between different unique IDs across different data sources

- **Personal details data:** the table gives demographic information for an individual across data collections in the IDI.

The following table gives the timespans and locations of each dataset within IDI:

Name		Timespan (tax period)	Location	Note
IR3	IR3 Key Points	1995-03-31 ~ latest (e.g., 2024-03-31)	IDI_Clean_YYYYMM: ir_clean.ird_rtns_keypoints_ir3	Before 2019 or 2020, IR3 Keypoints are not complete, it only covers people with self employment income
	IR3 Full Dataset	2000-03-31 ~ 2014-03-31 or 2013-03-31 ~ 2021-03-31	IDI_Adhoc clean_read_ir.ir_ir3_2000_to_2014 clean_read_ir.ir_ir3_2013_to_2020	These are addhoc dataset
IR Autocalc information:		2019-03-31 ~ latest (e.g., 2024-03-31)	IDI_Clean_YYYYMM: ir_clean.ird_autocalc_information	It does not contains the IR3 information, and it becomes more complete since 2019 or so (as more income components are added)
IR Personal Tax Summary (PTS):		2000-03-31 ~ 2020-03-31	IDI_Clean_YYYYMM: ir_clean.ird_pts	
Income by tax year summary		1995 – latest (e.g., 2023)	IDI_Clean_YYYYMM: data.income_tax_yr_summary	Largely some PAYE info, plus some self employment income and rent (not sure why Stats structured data like this)

Concordance	NA	security.concordance	
Personal details data	NA	data.personal_detail	

2. Procedures:

2.1 Obtain IR3 data (2000 - 2022)

2.1.1 IR3 (2013 - 2014)

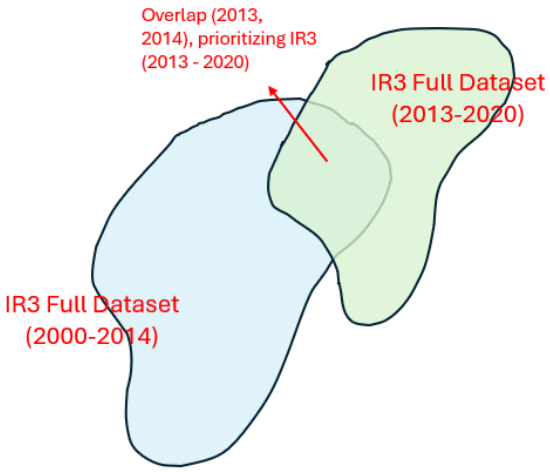
As previously mentioned, the IR3 “Full dataset” has two streams “IR3 2000 - 2014” and “IR3 2013 - 2020” within IDI, with overlapping data for the tax years 2013 and 2014. If multiple records exist for the same individual (identified by “snz_ird_uid” in the IDI) and tax period (identified by “period” or similar in the IDI), select the most recent record available.

Our goal is to isolate the IR3 data for 2013–2014 from this overlapping period, prioritizing the “IR3 2013 - 2020” dataset as it is the most recent.

Here’s a refined process to achieve this:

- Step 1: Extract data for 2013 and 2014
    - Extract data from both “IR3 2000 - 2014” and “IR3 2013 - 2020” for the tax years 2013 and 2014, using the tax return dates “31 Mar 2013” and “31 Mar 2014.”
- Step 2: Identify unique data in “IR3 2013 - 2020”
    - Isolate data that exists exclusively in “IR3 2013 - 2020” by filtering based on the individual unique ID (“snz_ird_uid” in IDI) and the tax period (“period” in IDI).
- Step 3: Identify overlapping data and prioritizing the data from “IR 2013 -2020”
    - Identify data present in both “IR3 2013 - 2020” and “IR3 2000 - 2014” using the same unique ID and tax period criteria. Remove the data from “IR3 2000 - 2014” if it exists in “IR3 2013 - 2020”
- Step 4: Combine data from Step 2 and Step 3
    - Combine the data from Step 2 (exclusive to “IR3 2013 - 2020”) and Step 3 (overlapping data prioritizing “IR3 2013 - 2020”).
- Step 5: Identify unique data in “IR3 2000 - 2014”
    - Extract data that exists only in “IR3 2000 - 2014” and not in “IR3 2013 - 2020.”
- Step 6: Handle and label the 2013–2014 Data
    - For the final 2013–2014 dataset, use the combined data from Step 4 and 5:
        - ✓ Data from Step 4 is labelled as “adhoc 2013 - 2020.”
        - ✓ Data from Step 5 (unique to “IR3 2000 - 2014”) is labelled as “adhoc 2000 - 2014.”

The following figures illustrate how the IR3 data between 2015 and 2018 is constructed:

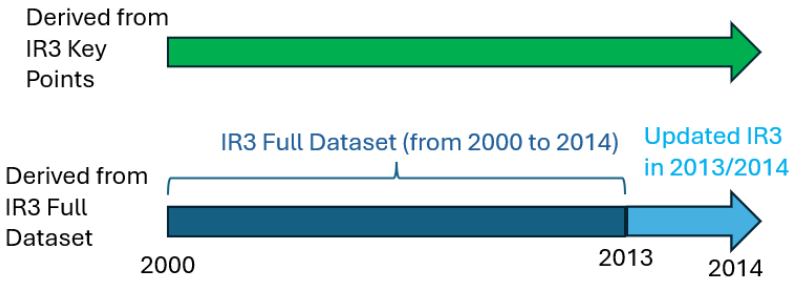


2.1.2 IR3 (2000 - 2014)

The IR3 data for the period 2000 to 2014 is derived from the following datasets:

- IR3 Full Dataset (2000–2012): Extracted from the broader IR3 Full Dataset (2000–2014).
- IR3 Updated 2013/2014 Dataset: Obtained from Step 2.1.1.
- IR3 Key Points (2000–2014)

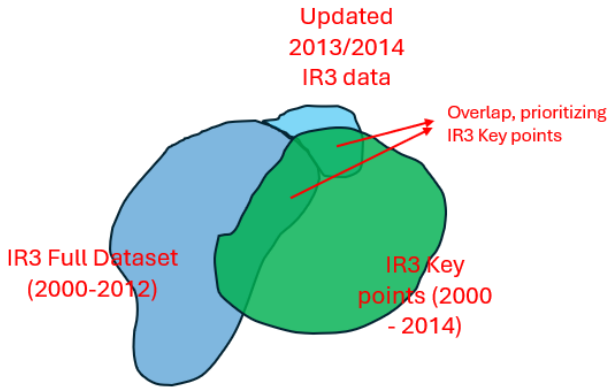
The process to compile this data is outlined below:



- Step 1: Extract 2000–2012 Data:
  - Retrieve data for 2000 to 2012 from the IR3 Full Dataset (the adhoc 2000–2014 data).
  - Note: If multiple records exist for the same individual and tax period, retain only the most recent record.
- Step 2: Combine 2000–2012 and 2013/2014 data
  - Merge the IR3 Full Dataset (2000–2012) from Step 1 with the IR3 Updated 2013/2014 Dataset (from Step 2.1) to create a combined dataset covering 2000–2014.
- Step 3: Extract IR3 Key Points (2000–2014):
  - Obtain the IR3 Key Points dataset for the period 2000 to 2014.
- Step 4: Identify data unique data:
  - Unique data to IR3 Key Points: Extract records present in IR3 Key Points (2000–2014) but absent from the combined dataset (Step 2), based on individual unique ID and tax period.

    - Identify data unique to IR3 Full Dataset: Extract records present in the IR3 Full Dataset (Step 2) but not in the corresponding IR3 Key Points (Step 3), based on individual unique ID and tax period.
- Step 5: Identify overlapping data
    - Determine records that overlap between the combined dataset (Step 2) and IR3 Key Points (Step 3), based on individual unique ID and tax period. Prioritize data from IR3 Key Points for these records.
- Step 6: Final compilation
    - Combine the following into a single dataset:
        - Unique data from IR3 Key Points (Step 4).
        - Unique data from IR3 Full Dataset (Step 4).
        - Overlapping data (Step 5), using IR3 Key Points as the primary source where applicable.

This results in a comprehensive IR3 dataset for 2000–2014, incorporating the updated 2013/2014 data. The following figure illustrate how the data is created:

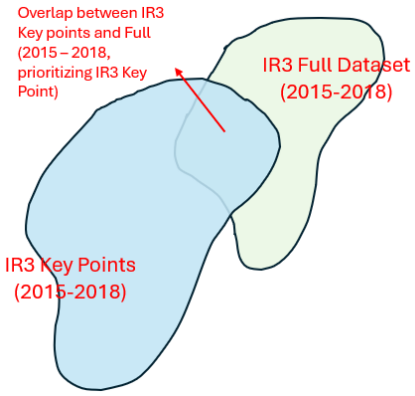


2.1.3 IR3 (2015 - 2018)

- Step 1: Extract Data (2015–2018)
    - From “IR3 Full Dataset (the adhoc one between 2013 and 2020)”: Retrieve data for the years 2015 to 2018. If multiple records exist for the same individual (identified by “snz_ird_uid” in the IDI) and tax period (identified by “period” in the IDI), select the most recent record available.
    - From “IR3 Key Points”: Similarly, retrieve data for 2015 to 2018, keeping only the latest record per individual and tax period.
- Step 2: Identify Overlaps, prioritizing the data from “IR3 Key Points”
    - Find Overlaps: Compare “IR3 Full Dataset” and “IR3 Key Points” to identify records that match based on “snz_ird_uid” and “period”. Remove the data from “IR3 Full Dataset” if it exists in “IR3 Key Points”
- Step 3: Identify unique data:
    - Unique Data in “IR3 Full Dataset”: Extract records present in “IR3 Full Dataset” but not in “IR3 Key Points.”
    - Unique Data in “IR3 Key Points”: Extract records present in “IR3 Key Points” but not in “IR3 Full Dataset.”
- Step 4: Merge the Data

- Combine the overlapping records (prioritizing “IR3 Key Points”), unique records from “IR3 Full Dataset,” and unique records from “IR3 Key Points” to create a more comprehensive dataset.

The following figures illustrate how the IR3 data between 2015 and 2018 is constructed:

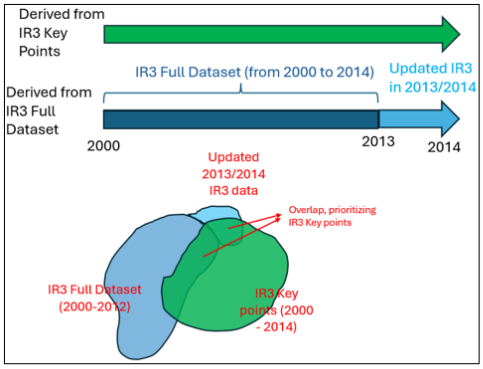


2.1.4 IR3 (2000 - 2022)

The IR3 data spanning 2000 to 2022 is compiled by integrating the following sources:

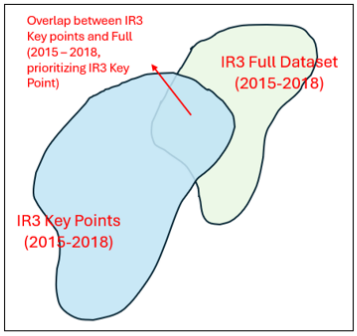
- Section 2.1.2: Derived IR3 data covering 2000 to 2014.
- Section 2.1.3: IR3 data covering 2015 to 2018.
- IR3 Key Points Dataset: Data covering 2019 to 2022.

The figure below illustrates the data sources used to construct the IR3 dataset for the period 2000 to 2022.



Derived IR3 data between 2000 and 2014

+



Derived IR3 data between 2015 and 2018

+

IR3 Key Points (2019 - 2022)

2.2 IR Autocalc information (2019 - 2022)

The IR Auto calculation dataset, spanning from 2019 to 2022, is sourced directly from IDI. If there are multiple records for the same individual (identified by "snz_id") and tax return period (identified by "ir_ac_return_period_date" in IDI), the most recent record is used.

2.3 IR Personal Tax Summary (2000 - 2018)

Similar to IR Auto calculation dataset, IR PTS between the year of 2000 and 2018 can be directly obtained from IDI. The most recent record is taken if multiple records available for the same individual and tax period.

2.4 Combine IR3, Autocalc and PTS data (2000 - 2022)

Based on Sections 2.1, 2.2, and 2.3, we have compiled the following datasets:

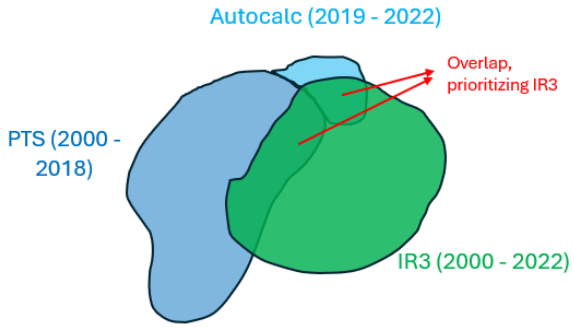
Data name	Period
IR3	2000 – 2022
Autocalc	2019 – 2022
PTS	2000 – 2018

To combine these datasets, follow these steps:

1. **Combine autocalc and PTS**: Merge the Autocalc and PTS datasets to create a joint dataset covering the period from 2000 to 2022, referred to as Autocalc + PTS (2000 - 2022).
2. **Identify unique records**: Determine the records that are unique to the Autocalc + PTS (2000 - 2022) dataset and not present in the IR3 (2000 - 2022), based on the tax period and individual unique ID.
3. **Merge with IR3**: Combine the IR3 (2000 - 2022) dataset with the unique records identified in Step 2.

Note that the Step 2 and 3 is a bit like combining Autocalc + PTS (2000 - 2022) with IR3 (2000 - 2022). For the overlapping parts, we give priority to IR3.

The following figure illustrates the data merging process.

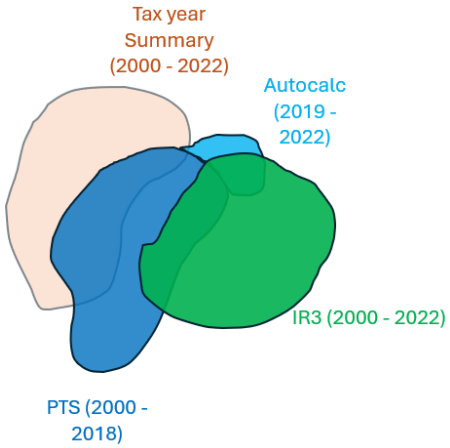


2.5 Incorporate Income Tax Year Summary data (2000 - 2022)

The Income Tax Year Summary dataset, available from 2000 to 2022, includes additional information not found in the PTS, Autocalc, and IR3 datasets.

First, we need to identify the unique entries within the Tax Year Summary data compared to the dataset from Section 2.4 (IR3 + Autocalc + PTS). Then, we can add this additional information to the IR3 + Autocalc + PTS dataset.

The following figure illustrates the data merging process.



This combined dataset contains the final income tax information between 2000 and 2022. The data later can be linked to other dataset such as personal details.

3. Output:

To save space, the data is yearly based on the tax return period.

The key output includes:

- Taxable income:
    - If the data source is “tax summary (Income by tax year summary)”, we take the data from “inc_tax_yr_sum_all_srces_tot_amt” (the total earnings for the individual for the tax year)
    - If the data source is “AC”, we take the data from “ir_taxable_inc_amt”
    - Otherwise take the data from “ir_ir3_taxable_income_amt” (the amount of taxable income for the period)

4. Summary:

This process involves merging various datasets to provide a comprehensive view of individual income tax from 2000 to 2022. In cases of overlapping data, the dataset with higher priority is used.

The IR3 dataset holds the highest priority. If there is overlap with other datasets (such as PTS, Autocalc, or the Income Tax Year Summary), the IR3 data takes precedence. The IR3 dataset is divided into three streams: IR3 Full Dataset (2000–2014), IR3 Full Dataset (2013–2020), and IR3 Key Points (2000–2022). Among these, the IR3 Key Points has the highest priority. For the IR3 Full Dataset, in overlapping years (e.g., 2013 and 2014), the latest data from 2013–2020 is prioritized.

The PTS (2000–2018) and Autocalc (2019–2022) datasets provide additional information to the IR3 data. However, in cases of overlap, the IR3 data is prioritized.

Similarly, the Income Tax Year Summary data (2000–2022) offers supplementary information to the combined PTS, Autocalc, and IR3 datasets. In instances of overlap, the combined PTS, Autocalc, and IR3 data is given priority.

The priority ranking for overlapping information is as follows:

- IR3 Key Points (2000–2022)
- IR3 Full Dataset (2013–2020)
- IR3 Full Dataset (2000–2014)
- PTS (2000–2018) and Autocalc (2019–2022)
- Income Tax Year Summary (2000–2022)