IN-CONFIDENCE

Ministry of Social Development

Social Outcomes Modelling 2023 – Technical Report

9 July 2024

Document classification: Client use

Wellington

Level 6, 22 The Terrace
Wellington 6011
P +64 4 974 5565

Sydney

Level 22, 45 Clarence Street
Sydney NSW 2000
P +61 2 9249 2900
P +61 2 9249 2999

Melbourne

Level 27, 459 Collins Street
Melbourne VIC 3000
P +61 3 9658 2333
P +61 3 9658 2344

ACN 087 047 809
ABN 29 087 047 809
www.taylorfry.com.au

Taylor Fry Pty Ltd

ISO 27001 INFO SEC Certified System

Climate Conscious
Certified carbon offsets
by South Pole

Table of contents

A

Introduction

1 Introduction to technical report

This is the technical report for the Ministry of Social Development's (MSD) modelling of social outcomes for New Zealand resident adults (aged 16+). The purpose of this report is to describe how the modelling project has been executed from a technical point of view. It serves as a record for reference and to allow consideration of the appropriateness of the approach taken. It includes details of the modelling construct, projection approach, assumptions made and data preparation.

The intended audience is anybody who wishes to understand how the modelling has been constructed and performed. The content is technical in places and is aimed at other modellers and technical analysts. We have included a high-level summary section which explains key aspects of the technical design which non-technical readers may find helpful.

The report is structured into the following sections:

- **Section 2 - High level modelling summary** – A layperson's description of the model. What it does, what it doesn't do, what it covers, and how it works.
- **Section 3 - Scope** – The scope of the model including the population covered and the services and outcomes that are projected. We also discuss compliance with actuarial and accounting professional standards.
- **Section 4 - Modelling approach** – A description of the modelling objectives and how the modelling approach has been designed to achieve these objectives.
- **Section 5 - Data** – A description of the data used in the modelling, including:
  - data sources
  - data checking performed
  - how the data was prepared for the quarterly projection process.
- **Section 6 - Assumptions** – The key assumptions made in the production of the datasets and model, and how these affect the model.
- **Section 7 – Projection** – A description of the projection process, including the projection components (population, outcomes etc.), the ordering of projection steps and the output of results.
- **Section 8 – Population model** – A description of how the population to be modelled is defined and how it is projected forward in time. This includes consideration of mortality, migration and children ageing into the adult population.
- **Section 9 – Benefit system models** – A detailed description of the models relating to future benefit use. This includes the component transitions that describe the movement of people from one benefit state to another and the associated benefit payments.
- **Section 10 – Public Housing models** – A detailed description of the models relating to future public housing use. This includes the component models relating to housing status, register status, market rent and Income Related Rent Subsidy. It also describes the process by which households on the register are modelled to enter into public housing (or otherwise exit the register).
- **Section 11 – Income models** – A detailed description of the models relating to income. This includes personal income, Working for Families (WFF) tax credits and NZ Superannuation.
- **Section 12 – Other Models**– A description of the modelling domains outside of the benefit, housing and income models. This includes Care and Protection and Youth Justice, Education, Justice and Health outcomes.
- **Section 13 - Quality assurance** – A description of the data and model checks and validation steps applied as part of our internal review process.

- **Section 14 - Reliances and limitations** – An articulation of risks inherent in the data and modelling approach and associated known limitations, including reference to any external inputs that we have relied upon.
- **Appendices** – A glossary of terms and acronyms, methodology for projecting regional unemployment rates, Generalised Linear Modelling theory and model coefficients.

2 High level modelling summary

This section gives an overview of the model in non-technical terms, answering core questions:

- What is the model?
- What does the model do?
- What outcomes does the model project?
- How does the model work?
- What does the model not do?

Section 3 onwards of this report captures the technical construct of the modelling and supporting data in a reasonable amount of detail. It is aimed at modellers and technical analysts, and so may be less accessible to other interested readers.

2.1 What is the model?

The term 'model' is broadly used to describe physical, mathematical and conceptual models. This model is a mathematical model. Many definitions of a 'mathematical model' centre on the notion of imitation or simulation i.e. a model imitates or simulates a real-world situation, often in a simplified way because the 'situation' being modelled is complex. In this sense, a model (including this one) might be described as a 'simplification of reality'.

Key aspects of the modelling framework for this project are:

- The population being modelled – In this case, New Zealand (NZ) residents aged 16 or older, and people entering this population over the next ten years.
- The future outcomes that are being modelled – See section 2.3 – What outcomes does the model project?
- The time horizon over which the future outcomes are being modelled – In this case, people's future lifetime.
- The historical data – Used to understand the correlative relationships between variables (or combinations of variables) and the future outcomes being modelled. Variables may be characteristics (e.g. demographics), relate to events (e.g. experience of the modelled outcomes in the past) or be environmental (e.g. measures of labour market conditions). Understanding the correlative relationships informs the construction of the mathematical equations that define the model, and the parameters for these equations.
- Assumptions – The model is underpinned by a range of assumptions which are either implied by the construction and parameterisation of the mathematical equations, or explicitly made. Explicit assumptions relate to variables that the model does not project but are built into model because they are important to projecting future outcomes, e.g. the future unemployment rate as a measure of future labour market conditions.

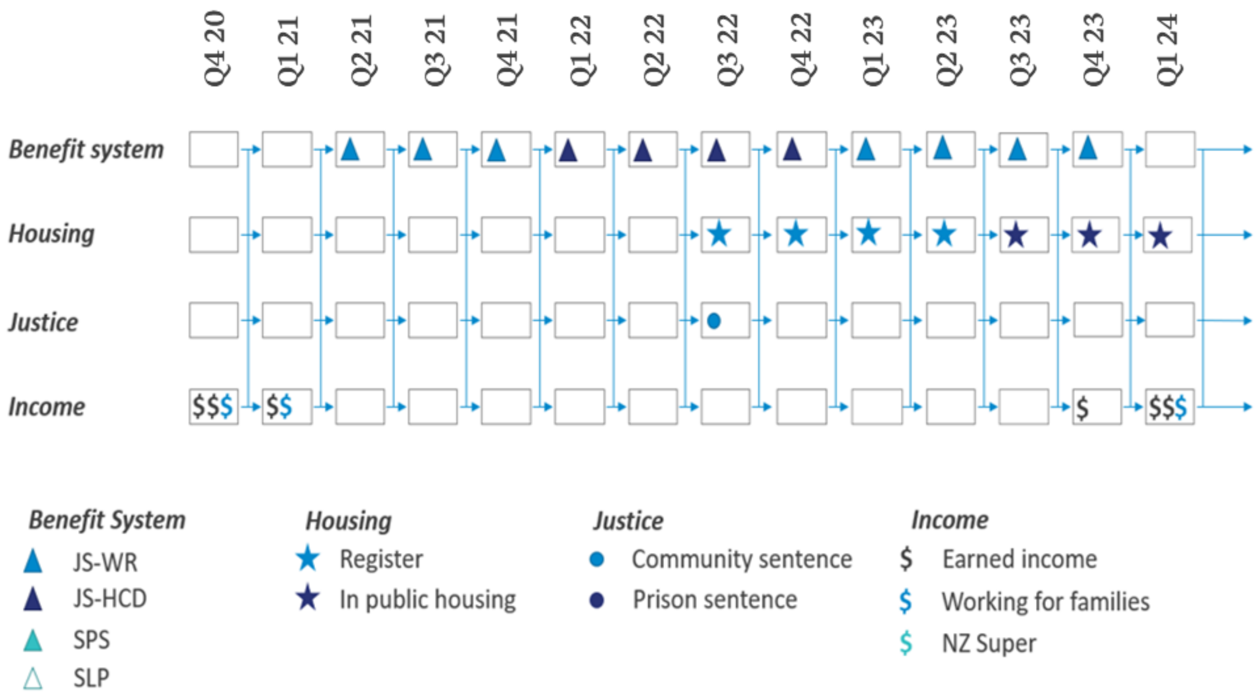
2.2 What does the model do?

In section 2.1 we referred to the model as a projection of future outcomes for a defined population (over 16-year-old NZ residents) over a defined time horizon (people's lifetimes). It does this by projecting people's status in relation to these outcomes (and other associated characteristics and outcomes) over each quarter-year period in the future. This is indicatively shown in Figure 2.1 below:

- For one person – a full model run produces similar projections for all NZ residents aged 16 and over.

- Over 14 quarters – a full model runs covers all people's future lifetimes and so runs for about 400 quarters.
- In respect of four outcomes – other outcomes are estimated by the model.

Figure 2.1 – Projected pathways



Where relevant, estimated cash flows are modelled in relation to future estimated outcomes. For example, benefit payments are modelled for those in receipt of a benefit and income related rent subsidies paid to public housing providers are modelled for people in public housing.

In addition to projecting outcomes for the present NZ adult population, the model also projects outcomes for those entering the population over the next 10 years. Population entry may happen in two ways:

- Ageing-in: children are considered to enter the adult population in the quarter in which they turn 16. We use projected output from the 2022 Oranga Tamariki children's model for this purpose.
- Migration: Both children and adults may enter the population via migration (which includes returning New Zealanders as well as foreign nationals).

Once in the population, outcomes for new entrants are estimated in the same manner as those in the present population.

2.3 What outcomes does the model project?

The model projects a large range of outcomes:

- **Benefit receipt** – This covers the incidence of benefit receipt and the associated payments. Benefit receipt is categorised into main benefit categories and supplementary assistance.
- **Other benefit receipt characteristics** – These include, but are not limited to partnered status, existence and age of children, and incapacity coding for health-related benefits.
- **Public housing** – This covers entry to the public housing register and associated prioritisation rating, movement off the register (either into public housing or otherwise), income related rent subsidy, exit from public housing, size and location of house allocated, and future dissolution of households currently in public housing.

- **Income** – This covers personal income, Working for Families (WFF) tax credits and NZ superannuation. The primary industry from which personal income is earned is also modelled.
- **Justice activity** – This covers number and type of offences committed as well as community and custodial sentences managed by the Department of Corrections.
- **Education** – This covers secondary and tertiary enrolment in the quarter, secondary attainment, total days of any suspensions or stand downs at secondary school, highest New Zealand Qualification Framework (NZQF) level enrolled and attained at tertiary.
- **Child and protection (CNP) and Youth Justice (YJ)** – This covers the highest level of either type of intervention as well as the total number of days spent in placements.
- **Health** – the covers mental health and addiction pharmaceutical, specialist community and specialist inpatient events, acute hospital discharges and mortality.
- **Location** – this covers the region/TLA/Auckland board where an individual resides.

2.4 How does the model work?

Figure 2.1 highlighted how the model projects outcomes at each quarterly time step.

Referring to the model as a 'model', implies that it is single model. In fact, it is made up of over 200 individual models. Each of these individual models plays a specific part in the overall modelling construct. Some relate to how a person moves between different outcome states from one quarter to the next e.g. benefit state. Some relate to the evolution of other modelled outcomes e.g. personal income. Others relate to cash flows associated with particular outcomes e.g. benefit payment given an individual is projected to be receiving a benefit in a quarter.

The vast majority of the models fall into the broad category of models known as regression models, which means they estimate one variable based on other variables. The remainder of the models are probability table models that attach probabilities to different outcomes.

The models are pulled together in what we refer to as the 'projection code'. Many of the variables that each individual model relies upon are themselves modelled variables. For example, the models relating to transitioning between benefit states from one quarter to the next depend on, say, corrections activity variables which, in turn, are updated each quarter. The projection code runs each model in a set sequence for a future quarter, before moving onto the next quarter and repeating the sequence based on the updated variables. For this reason, the overall modelling construct is sometimes referred to as a 'chained regression model': it chains together regression models over a series of future time steps (in this case quarters).

2.5 What does the model not do?

The model is not a causal inference model. By this, we mean that the model does not attempt to determine the causal factors relating to different outcomes. Rather, the model is a predictive model, and thus seeks to determine factors that are correlated with outcomes. This difference is important. For example, a key finding of previous work is that long-term dependence on welfare is highly correlated with those who first receive benefits when under twenty years of age. So, age of first benefit is highly predictive of lengthy spells supported by benefit. However, it cannot be concluded that this is the cause of these spells. Nevertheless, knowledge about correlations and relationships between certain characteristics and outcomes is valuable information for policy and programme design and monitoring.

The model is based on simulating individual pathways through various welfare and housing states (including not receiving any benefit/assistance) as well as other characteristics (family information, education, income, corrections sentences etc) over their lifetimes. There are many possible pathways from the modelling projection date to time of death, so the exact pathway is very uncertain. Results for any particular individual reflect the average for people with similar characteristics and are not intended to be an accurate prediction of that individual person's future pathway. Results, therefore, should be considered for segments of the population, rather than at an individual level.

3 Scope

This section describes the scope of the modelling. Specifically:

- The population for which outcomes are projected
- The outcomes and associated cashflows being projected
- The time horizon over which outcomes are projected.

3.1 Population

The population in scope is all people aged over 16 years old[1] and resident in New Zealand as at 30 September 2023. Over the first 10 years of the projection, additional people enter the population to allow for children ageing into the adult population and to reflect the effects of positive net migration into New Zealand. The projection runs from the effective date (30 September 2023) until projected death of all individuals covered in the population defined above.

3.2 Outcomes and associated cash flows

Included in the scope is a range of outcomes:

- Working-age benefit receipt
- Public housing register applications
- Public housing placements
- Personal income, WFF tax credits and NZ Super
- Industry employed in
- Corrections and offence activity
- Educational outcomes
- Child protection and Youth Justice outcomes
- Mental health, mortality and hospitalisation events.

Table 3.1 below provides more detail on these.

[1] While many definitions of 'adult' exist, we commonly refer to those aged over 16 as adults throughout this report.

Table 3.1 – Outcomes and associated cash flows in scope

Category	Outcomes modelled	Associated cash flows modelled
Benefit payment (tier 1)	▪ Jobseeker Support ▪ Sole Parent Support ▪ Emergency Benefit ▪ Youth Payment ▪ Young Parent Payment ▪ Supported Living Payment ▪ Orphan's/Unsupported Child's Benefit	Yes
Benefit payment (tier 2)	▪ Accommodation Supplement ▪ Disability Allowance ▪ Child Disability Allowance ▪ Winter Energy Payment	Yes
Benefit payment (tier 3)	▪ Hardship Payments including Temporary Additional Support ▪ Recoverable assistance	Yes
Benefit payment (other)	▪ Childcare subsidy ▪ Employment interventions	Yes
Other benefit characteristics	▪ Partnered status ▪ Number of dependent children and age of youngest child (main benefit clients only) ▪ Incapacity code for JS-HCD and SLP ▪ Proportion of quarter supported by main benefit	N/A

Category	Outcomes modelled	Associated cash flows modelled
Public housing	▪ Tenancies in Kāinga Ora or Community Housing Provider (CHP) managed properties and associated Income Related Rent Subsidy (IRRS) costs ▪ Applications to the public housing register ▪ Transfer applications (both client and business initiated) ▪ Accommodation Supplement[2] ▪ Temporary Additional Support[3] ▪ Emergency Housing	Yes
Income	▪ Personal income ▪ Working for Families tax credits ▪ NZ Superannuation payments ▪ Industry in which employed for earned income	Yes
Justice	▪ Percentage of time serving any sentence (community or custodial) over the last quarter, excluding driving-related offences ▪ Percentage of time serving a custodial sentence over the last quarter ▪ Percentage of time serving any sentence in the last quarter relating to a theft offence. ▪ Number of times proceeded against by Police in the quarter, and category of alleged offence that led to the Police proceeding	No
Education	▪ Whether the person has left secondary school ▪ The attainment level at secondary school ▪ The total combined days of all suspensions or stand-downs while at school ▪ Whether the person is enrolled in tertiary education ▪ The highest NZQF level of tertiary enrolment to date ▪ The highest NZQF level of tertiary completion to date	No

[2] Accommodation Supplement is a benefit payment which is largely related to housing costs. It bridges the two systems so we have listed it twice, but we do not double count this item.

[3] Similar to above, Temporary Additional Support is a hardship benefit payment, which is largely used to cover housing costs. Again, we have listed this twice, but do not double count this item.

Category	Outcomes modelled	Associated cash flows modelled
Child protection and Youth Justice	▪ Highest level of care and protection intervention to date ▪ Highest level of Youth Justice intervention to date ▪ Days in care and protection placement to date. ▪ Days in Youth Justice placement to date	No
Health events	▪ Mental health pharmaceutical events ▪ Mental health specialist community events ▪ Mental health specialist inpatient events ▪ Acute hospitalisation days ▪ Mortality	No
Location	▪ Region ▪ TLA/Auckland Board	N/A

3.3 Notable exclusions to scope

3.3.1 Exclusion of Jobseeker Support – Student Hardship

In previous years' modelling, it was judged that the Jobseeker Support – Student Hardship was not an appropriate benefit type to include in the projection for the following reasons:

- Most other financial assistance provided to students is excluded. The exception to this is Student Allowance, which has been included as part of income.
- The benefit is highly seasonal - students only receive the benefit if they cannot find employment in the summer holidays. This pattern is less amenable to management, as the concept of a long-term benefit system client is not applicable.
- The relationship between this benefit and other key benefits is uncertain and has the possibility of skewing the main transition models.

Therefore, client spells on this benefit have been ignored, both in terms of projecting cash flows and determining which clients are supported by a benefit.

We have retained this approach for the current model, although we recognise that the reasoning for exclusion may not always be relevant and, therefore, could be considered for inclusion in the scope for future years.

3.3.2 Other housing expenses

This report does not attempt to estimate other 'costs' associated with public housing, such as:

- Any charge, where applicable, for the cost of capital.
- Kāinga Ora (and other public housing provider's) administrative expenses, rates, or costs of repair. This is mainly to avoid double counting; the market rent of a property in the private market typically includes the cost of property management and maintenance by the landlord. However, these costs are

potentially important and may warrant future analysis; understanding how management costs vary across households is useful in understanding household-level need.

- Future costs for renewal and reconfiguration of the current public housing stock.
- Any measurement of 'unknown' demand (for example, the potential housing costs of people who would qualify for a public housing placement but currently do not apply).

3.3.3 Cross-sectoral costs

The projection includes models which project service usage across areas outside of MSD's purview. These include:

- Corrections activity and police proceedings
- Educational enrolment
- Child protection interactions for young adults
- Youth Justice interactions for young adults
- Mental health and substance abuse interactions, acute hospitalisations and mortality.

In theory costs could be attached to these estimates of outcomes and included with other lifetime cost estimates. This was deemed out of the scope for this projection for the following reasons:

- To focus attention on domains within the control of MSD. It is unclear how MSD would be expected to respond to results relating to population crime or child protection trends.
- To avoid duplication. There is other modelling which focusses on these services e.g. Oranga Tamariki's children's model.

3.4 Time span of projection

Modelled outcomes and related state variables are projected for each quarter from the effective date (30 September 2023) until the forecast death of all cohort members – around 100 years. Estimates of specific outcomes stop at ages consistent with the eligibility of the related services e.g. 65 for main benefit receipt.

3.5 Changes since previous report

None.

4 Modelling approach

In this section we give an overview of the modelling approach, laying the foundation for more detailed sections later in the report.

4.1 Introduction

An annual projection model of the benefit system has been a key part of MSD analytical tools since 2012, providing key measurements of progress as well as determining the factors that affect long-term trends and costs. In 2015, the modelling was extended to incorporate a model of the public housing system, with inter-dependencies between welfare and housing utilisation.

The present model has a wider scope still, focussing on the social outcomes of the New Zealand adult resident population. Earlier years' reporting has focussed on the results of modelling, with less detail on the technical specification. This is the sixth year we have provided a report specifically focussed on the technical specification. The key purposes of the report are to:

- Ensure a common understanding of the technical specification of the modelling
- Describe the modelling process we have been through
- Ensure a common understanding of what the model covers (and by extension what it does not cover)
- Provide a basis for considering the applicability of the technical specification in meeting the objectives of the modelling
- Provide a reference basis from which modellers of equivalent capability could replicate the projection code if needed.

4.2 Background

Figure 4.1 provides a high-level view of the projection process and its various components: population, models and assumptions.

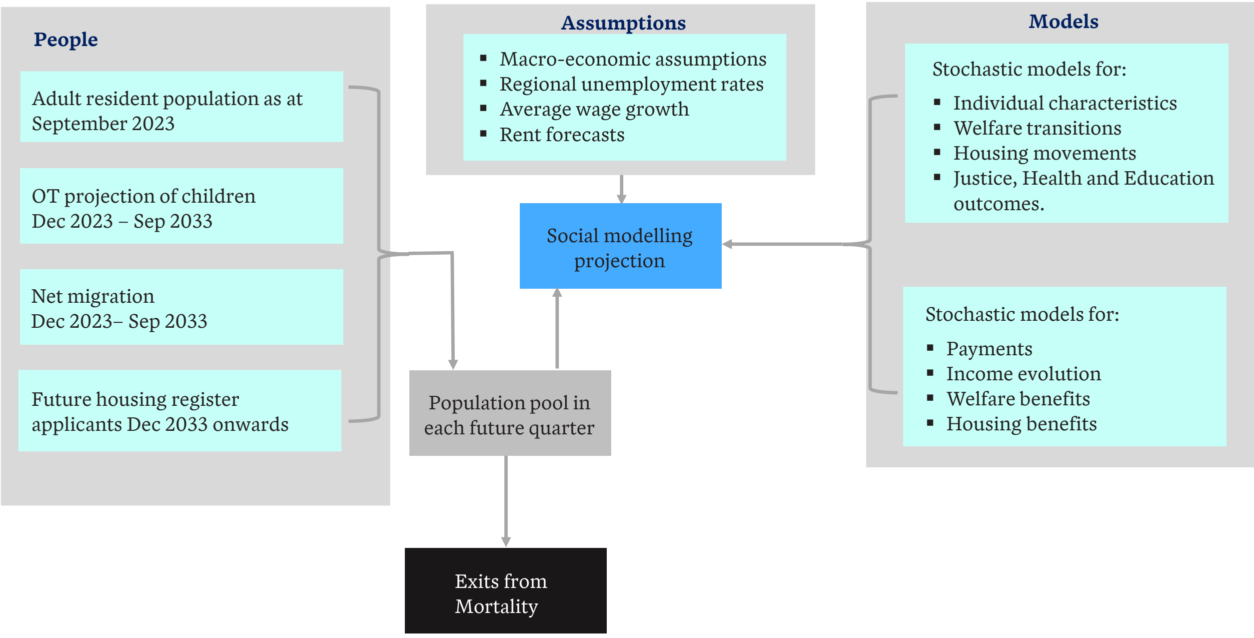
The projection focusses on the current New Zealand adult resident population and on the projected population over the next ten years. It projects various social outcomes to including:

- Lifetime welfare and public housing utilisation and costs
- Income levels, and Working for Families tax credits
- Employment and industry.

The projection is carried out on an individual, quarterly basis from the projection date (or date of entry into the adult population) until death. Estimates of lifetime support vary based on factors including:

- Sociodemographic factors (age, gender, region, prioritised ethnic group, educational attainment)
- Macroeconomic factors (regional unemployment rates and housing market rental rates)
- Past benefit receipt
- Past housing support history
- Family benefit history ('intergenerational') variables
- Past child protection and Youth Justice interactions
- Past criminal convictions
- Past mental health and hospitalisation events.

Figure 4.1 – Overview of the projection



In all, approximately 200 models, which make use of hundreds of variables, are used to project extremely detailed joint pathways through the housing and welfare systems and projections of other outcomes.

4.3 Model design

The current model has its origins in the original baseline valuation of the New Zealand welfare system. In subsequent years, it has been extended to incorporate additional population characteristics and public housing.

The model itself is designed as a micro-simulation of the New Zealand population. By this, we mean that we model people individually and consider their evolution over time. However, there are many possible pathways from the modelling date to time of death, so the results for any particular individual are reflective of the average for people like them rather than being credible predictions specific to that person. Results, therefore, should be considered for segments of the population, rather than at an individual level.

A brief overview of the key design components is given here; later sections provide more detail.

4.3.1 Population

The model is applicable to the full New Zealand adult resident population over the next ten years. Therefore, as well as the population as at 30 September 2023, new entrants (children ageing in, positive net migration) must also be incorporated at an individual level. Section 8.3 contains further details on the processes used for this.

4.3.2 Individual characteristics

Each individual has a number of key characteristics attached to them which are used as predictors in the various micro models that make up the global model. These characteristics include things such as age, sex, ethnic groupings (henceforth referred to as ethnicity), education, incapacities (if any), corrections history, etc. The model design separates these into characteristics that remain static (e.g. ethnicity, sex) and those that evolve over time.

4.3.3 Welfare

Welfare utilisation is both an important insight and an important predictor for public housing usage (and vice versa). Design-wise, the model makes the assumption that welfare clients access one benefit type per quarter. This is the main benefit type they receive the most, if any. If none, then a supplementary benefit type. It then considers transitions from one benefit type to the next. There are nine states in total for working-age individuals including 'Not on Benefit'. A series of payment micro-models completes the set of welfare sub-models.

4.3.4 Housing

There are three housing states – in a public house, receiving Accommodation Supplement (AS) and NIL (not in public housing or receiving AS). Note that being on the housing register is not a separate housing state as it is possible for individuals in all three housing states to be on the register. Therefore, in addition to housing state, we maintain a flag for those on the register.

A key design feature of the housing model is that it enforces proper supply and demand dynamics for housing – a house is only allocated when there is a free house, and a priority ordering of individuals on the housing register is used for this housing allocation.

Note that this requirement has major consequences for the design of the projection code, since it means that, unlike welfare, housing movements cannot be projected independently for each individual.

Models for market rents, IRRS, AS, TAS and emergency housing payments are used to determine future housing-related costs.

4.4 Considerations in model development

For a complex prediction tool such as the social outcomes model, there are several key decisions to make around the architecture, types of models used, prediction approaches, etc. These are outlined below.

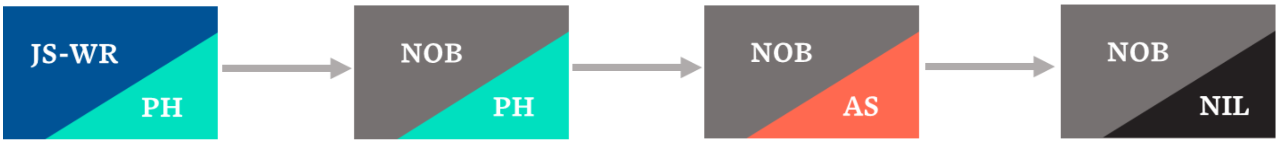
4.4.1 A single projection model

A key design decision in our 2015 report was to combine the benefit system and housing projection models. The substantial overlap between these systems means that considerable insight is gained from a combined approach. The current model is an extension of the prior housing and benefit system model to the full population, so it is natural to retain the single projection model approach.

4.4.2 Housing and Benefits

Households may be receiving both main benefits and a housing-related payment in any given quarter. Both are important for understanding and predicting household and benefit pathways. As a result, the core projection model projects dual welfare-housing status each quarter. For example, Figure 4.2 shows a pathway in which a primary tenant is receiving both Jobseeker Support – Work Ready benefit (JS-WR) and residing in public housing (PH) in the first quarter. In the second quarter, the primary tenant becomes employed and ceases to receive a main benefit (Not on Benefits - NOB) while still residing in public housing. In the third quarter, the household leaves public housing and receives Accommodation Supplement (AS). In the fourth quarter the household becomes fully independent, receiving neither a main benefit nor a public housing payment (NIL). Note that this is an idealised pathway demonstrating progressively greater independence in each state.

Figure 4.2 – Illustration of joint welfare benefit and housing quarterly transition states



4.4.3 General population

An important aspect of understanding future social outcomes of the general population is to estimate their future benefit system utilisation and public housing costs. Incorporating them as a sub-population into the existing model is the logical, and indeed necessary, step in the case of public housing where limited supply means that all housing applicants must be considered in tandem to take into account supply and demand dynamics.

4.4.4 Independence of individuals

A key technical issue is whether people are projected independently of others – it simplifies many things and reduces the computational burden but at the cost of losing some predictive information and insights. To strike a balance between ease of computation and retaining details, we have opted for an architecture where individuals are considered independent for the most part except that:

- Those starting in public housing are grouped together until they exit their current house. In particular, transitions out of public housing for those in the same house are correlated.
- Proper public housing supply constraints are enforced – houses are only allocated once they have been vacated.

4.4.5 Transition model approach

The transition model approach focusses on understanding how people move through the benefit and housing systems over time. It is worth mentioning here that there exist alternatives to such an approach (see for instance, the snapshot-based approaches used in Section 15 of the 2012 valuation report for the segmentation analysis). However, we have chosen the transition approach for several reasons:

- Responsiveness: Changes in movement behaviour observed in recent years can be correctly reflected in the models.
- Long range prediction: We can leverage the behaviour of clients at various stages of the benefit system to make long-range assumptions. For instance, we assume that the behaviour of older clients may be used to model the behaviour of the younger clients in the distant future. Given this assumption, the transition approach enables us to forecast full future outcomes for younger clients.
- Intuitive appeal: A focus on measures such as probabilities of entering and exiting benefits is natural and allows easier drill-down analysis.
- Consistency: The approach worked well for the welfare scheme in both the first aggregate level (Level I) valuation and the segment level (Level II) valuations performed on 2011 and 2012 data. Since then, a similar approach has been used in a further six annual projections of cost and utilisation. The integrated housing and welfare projection model was first applied to the June 2015 cohort and has been updated every year since. Therefore, the approach may be considered mature with a well-established track record.

The same transition approach is applied to the general population (i.e. those who have never been on welfare or in public housing) to predict their future pathways through the two support systems. Additionally, the income modelling, although not a transition model per se, makes use of similar ideas in that past income levels are key predictors of current levels.

4.4.6 Component models

The modelling process may be considered in a hierarchy where key processes are modelled using detailed stochastic models and lower-impact factors are modelled using simpler models. The specific types of models used in the projection include:

- Generalised Linear Models (GLMs): these are a powerful multivariate tool for modelling probabilities and amounts. They both provide useful information on key drivers and enable projection results to respond to changes in key drivers. Key transitions and payments and some key parts of individual characteristics evolution in the projection are modelled using detailed GLMs.

- Multinomial models: these are generalisations of a Binomial GLM to deal with multiple outcomes. The model estimates the probability of each outcome.
- Other stochastic models: other models used include beta models, zero and one inflated beta models (these allow for a probability mass at zero/one, then distribute the remaining probability over a fixed range of numeric outcomes) and some normal and log normal models.
- Probability table models: these are used for a discrete number of outcomes where the table is in the form of a look-up table. The probabilities usually depend on a number of key characteristics. For example, a probability table is used to determine the incapacity of a new entrant to one of the disability related benefits (JS-HCD and SLP-HCD) from a non-disability state. These particular look-up table probabilities depend on age and sex.

In general, the transition models (movements between the different welfare states and housing states) are modelled using GLMs for the most important transitions (remaining on the same benefit, moving off benefits) with multinomial models to fill in the full transition matrix. Individual characteristics that evolve over time use a mixture of parametric statistical models (mostly GLMs) and non-parametric probability tables. Payment models are GLMs; this recognises the fact that while benefit levels may be close to deterministic, duration effects do lead to variation in payments which must be estimated.

4.4.7 Simulated versus exact projection

A key design choice was whether to calculate exact results based on the underlying models or to estimate approximate results using a simulation approach. The differences between the two are explained below:

- Exact: this approach tracks every possible outcome for each client for every future quarter and its associated probability based on the underlying models. This process has a heavy computational load due to the many possible outcomes for each client.
- Simulation: this approach follows each person through time, using the transition probabilities to simulate a single path for a client. This process is then repeated many times to determine many possible paths for each client. This is also computationally intensive, though less so than the exact approach unless a very large number of simulations are run.

In many ways, the exact approach is preferable; for instance, it gives more correct estimates of the mean, and on the relative likelihood of rarer events. This approach was taken in the 2011 projection model. Subsequently, the addition of extra benefit states and modelling variables made the exact approach computationally intractable. Therefore, we have adopted the simulation-based approach since the 2012 projection model.

4.4.8 Software

It is also important to consider the software as part of the model design since different functional capabilities may influence model details.

Prior to 2018, all modelling and projection work was carried out using SAS, with a customised projection combining the many models and assumptions into a cohesive whole to produce predictions and insights.

Since September 2018, the projection modelling work is based on microdata in the IDI and thus all the modelling and projections must be carried out using IDI hardware and software. Consequently, a review of the modelling and projection process was carried out before starting the September 2018 projection and it was decided to recode the projections in R. There were several reasons behind this decision:

- Computational: R has good parallel libraries, which means that there is potential to speed up the simulations by running calculations in parallel.

- Formal programming language: R is better suited to complex tasks, being a more formal programming language than SAS. It is particularly well suited to statistical programming and enables us to easily integrate the various stochastic models into a holistic projection framework. Furthermore, due to its popularity among statisticians and data scientists, it has an ever-growing rich set of libraries to improve its functionality.
- Vectorised language: R is a vectorised language, which means, in effect, that it is more efficient to do calculations for many individuals at the same time rather than loop through each individual separately. This feature is particularly useful when there are inter-dependencies between individuals (such as households in public housing) since it makes managing these dependencies more straightforward than in an individual observation–based paradigm such as SAS.

B

Data and assumptions

5 Data

Data is the foundation of any modelling project. In this section we discuss:

- The datasets we have used and the timeframes they cover
- The quality of the data
- Notable issues identified with the data
- The data preparation process.

5.1 Source data

The Integrated Data Infrastructure (IDI) is the central database for this project and is maintained by Statistics New Zealand. All data contained within the IDI are anonymised to protect the privacy of individuals and organisations. In particular, use of data is governed by the Data and Statistics Act 2022, Privacy Act 1993 and the Tax Administration Act 1994 which ensure the privacy of the data utilised. The main benefit of the IDI data is the cross pollination of data from different government sources, allowing the modelling process to accurately characterise individuals' needs for government services.

The data we have used for modelling is from the June 2023 IDI data refresh. However, for projection we have utilised monthly refreshes of the benefit system data and the Inland Revenue EMS table to enable us to run the simulation model from 30 September 2023. We have provided a reconciliation of the data refreshes in Section 5.3.2 below.

A summary of the data sets employed in the project is presented in the table below:

Table 5.1 – List of datasets utilised by the long-term modelling from the IDI

Data set source	Data set name	Description
Department of Corrections	Sentencing and remand	Data on convicted offenders who receive a community sentence or imprisonment, and people remanded until their trial is completed.
Department of Internal Affairs	Births	A record of each birth in New Zealand containing details of parents and child. Only used for determining the population.
Kāinga Ora	Tenancy data	Contains one record per household per month for those in public housing. Includes general household information such as income, income related rent subsidy, etc.
	Tenancy household data	Contains one record per person per month for those in public housing. Contains demographic information (although this is overwritten by personal details demographic information).
	Register data	Contains one record per application (i.e. per potential household) per month for those on the public housing register. Contains

Data set source	Data set name	Description
		information related to the application such as preferred location and priority scores.
	Register exit	Contains one record per application and the date when that application exited the register.
	Register transfers	Contains the same information as the register data but only if the application is from someone currently in public housing.
	Register applications	Contains the same information as the register data but only if the application is from someone not currently in public housing (i.e. not a transfer application)
	Register household data	Contains one record per person per month for those who are part of an application on the public housing register. Contains demographic information (although this is overwritten by personal details demographic information).
	Register status changes	Contains one record per application per change in register status. This is used to add back in applications that are on hold or have been offered a house but are not in that house yet.
	Houses data	Contains one record per month per Kāinga Ora administered house. Contains house characteristics such as location and number of bedrooms.
Immigration New Zealand	Visa decisions	Contains records on NZ visas applied for and received. Only used for determining the base population.
Inland Revenue	EMS	Employee-level data from employer monthly schedule filing requirements. Includes earnings, employer and deductions. Records are monthly.
	Returns Keypoints IR3	Contains information on individuals who have filled out an individual tax return (IR3). Contains partnership, self-employment and shareholder salary income. All information is recorded on a March year basis.
	Attachments IR20	Contains information on salary income on individuals from the IR20 return. This return is filled out by partnerships on the salary

Data set source	Data set name	Description
		income of partners. All information is recorded on a March year basis.
	Attachments IR4S	Contains information on salary information individuals from the IR4 return. This return is filled out by a company on the salary income of shareholders. All information is recorded on a March year basis.
	Customers	A spell file that contains information from Inland Revenue on individuals. We have used it to obtain industry information for those people who are not in the EMS.
	Regular refresh EMS	A dataset for each quarter in 2023 that includes similar information to the EMS.
	Spells (WFF)	Contains individual data regarding start date of a spell receiving a main benefit, duration and end date and amount received. Provides information on WFF tax credits, Accommodation Supplement and Childcare assistance.
Ministry of Education	Industry Training (tec_it_learner)	Contains data regarding education-related activity occurring in schools, tertiary education and workplace-based training settings. Only used for determining the starting population.
	Student Enrol	Collection of data that records a student's enrolment in a school and a student's termination of that enrolment.
	Student Leavers	Contains information on when an individual has left secondary school.
	Student interventions	Contains information on type and duration of interventions including student stand downs or suspensions.
	Student qualification	Records information on qualification achieved, attempted, date attended and education provider.
	Enrolment (tertiary)	Records information regarding course enrolments with tertiary education providers.
	Completions (tertiary)	Records information regarding course completions at tertiary education providers.
Ministry of Health	PRIMHD	The Ministry of Health collection of national mental health conditions and addiction

Data set source	Data set name	Description
		information on service activity and outcomes for healthcare users.
	Pharmaceuticals	Data containing claims and payment information from pharmacists and subsidised medication.
	Publicly funded hospital discharges	Records hospitalisation discharges and information on related conditions. From National Minimum Dataset (NMDS).
	Primary health organisation data	Records enrolment with primary health organisations. Used for determining the population only.
Ministry of Social Development	Spells	Contains information on spells individuals have spent on primary benefits.
	Partner	Contains information on partners while on a primary spell.
	Child	Contains information on any children while on primary spell.
	District	The MSD district that applies to each client for a spell.
	Incapacity	Further detail on what type and the number of incapacities those on health benefits have.
	SWN	Contains demographic information on MSD clients. Only used for the education level MSD has recorded for clients. Most of the other information has been sourced from the personal details table.
	Childcare subsidy (from Adhoc library)	Contains information on childcare subsidies paid. Does not contain backdated payments that were made.
	Supported Living Payment Reassessments (from Adhoc library)	Contains information on reassessment periods for clients receiving health benefits.
	First Tier Expenditure	Contains payment information for spells on first tier benefits.
	Second Tier Expenditure	Contains payment information for spells on second tier benefits.
	Third Tier Expenditure	Contains payment information for spells on third tier benefits.

Data set source	Data set name	Description
New Zealand Police	Police proceedings	Contains information on individuals proceeded against by the New Zealand police. Used for offence leading to police proceedings data.
Oranga Tamariki (previously Child, Youth and Family)	Intakes events and intakes details	Information on those who have had a report of concern but no further action.
	Investigations events and details	Information on care and protection investigations.
	Legal status	Information on youth high court events.
	Placements events and details	Information on the number of placements for an individual and their total number of days in placement.
	Family group conference	Information on those who have had a Youth Justice family group conference.
Statistics New Zealand derived data	Personal details table	Contains demographic information such as birth months, ethnicities and sex. Sourced from various agency datasets.
	Income tax year	Income amounts per person per month. Data is sourced from various Inland Revenue tax datasets. For some sources income is only recorded on a March year basis rather than monthly.
	Overseas spell	Records the spells an individual has been overseas. Used for checking the overseas rules in defining the base population.
	Address notification table	Contains a spell file of the historical address records for an individual. Sourced from various agency datasets that contain addresses.

5.2 Data gaps

The different data sources used each cover different data periods. The following diagrams summarise which periods are covered by each dataset used for modelling outcomes and each dataset used to define the population.

Figure 5.1 – Time period covered by datasets used to model outcomes

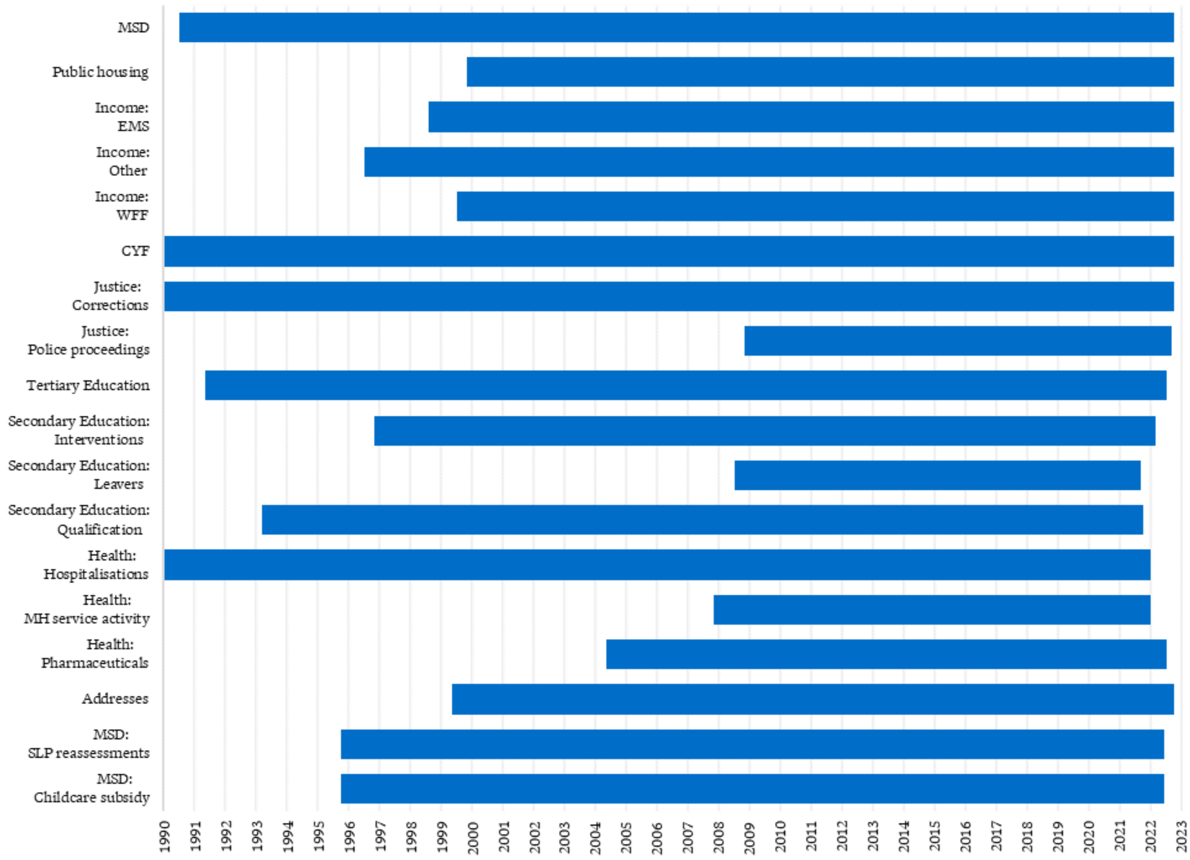
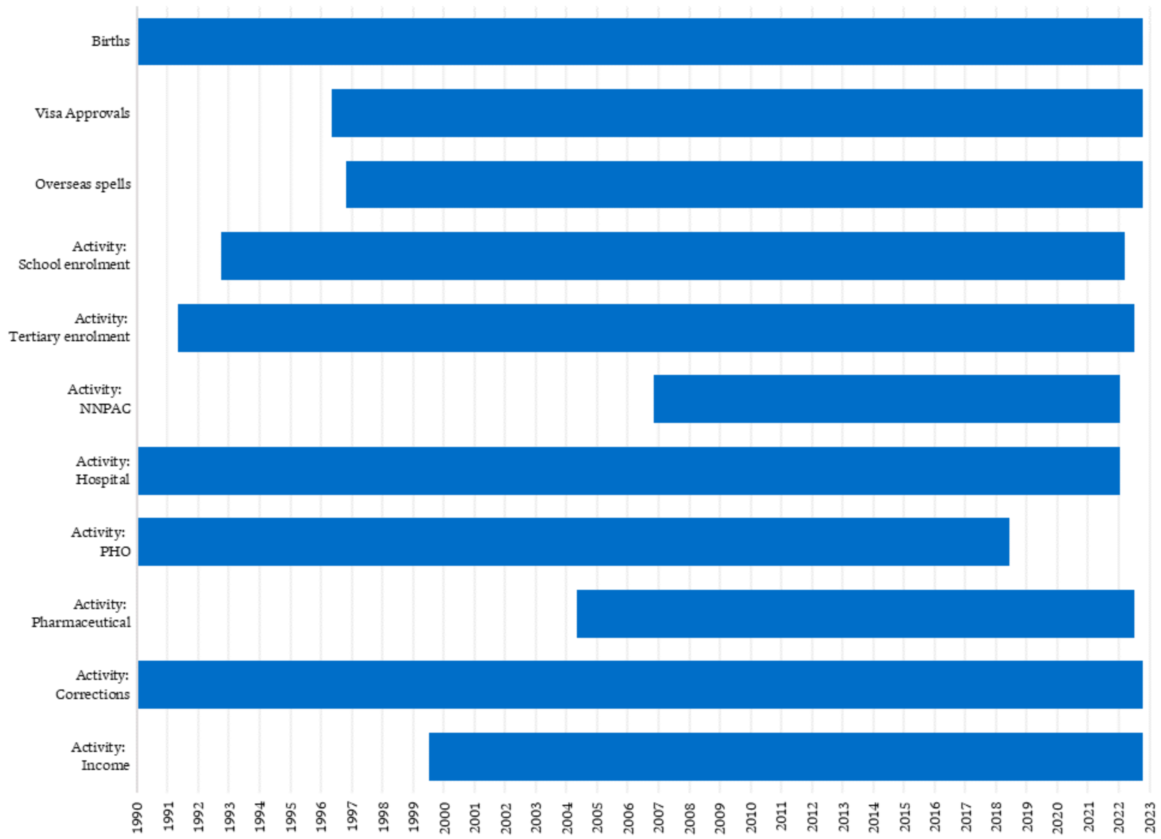


Figure 5.2 – Time period covered by the datasets used to define the population



Where data sources do not go to September 2023 they have been imputed to do so. More information on this is included in Section 5.3.6.

5.3 Data quality

5.3.1 Linking

Statistics NZ are responsible for linking data within the IDI from different agencies. Individual records are analysed across data sets to identify which records likely belong to the same identity. One aspect of the quality of an individual dataset in the IDI can be measured by match rates with the IDI spine. The IDI spine is the central dataset of individuals against which all datasets are linked. A dataset with a low match rate means that individuals in the dataset can't be linked with data from other sources.

The IDI spine is the primary person-level dataset that is considered the highest quality representation of the resident New Zealand population. Therefore, a good measure of data quality for a given data set is the proportion of records that match to the IDI spine[4]. Match rates vary across different data sets, a summary of which is given in Table 5.2.

Table 5.2 – Match rates to the IDI spine

Data set source	Data set name	Spine match rate
Corrections	Sentencing and remand	98%
Child, Youth and Family; subsequently Oranga Tamariki	Intakes & investigations events and details	93%
	Placements events and details	99%
	Legal Status	98%
	Family group conference	97%
Housing New Zealand subsequently Kāinga Ora	Tenancy data	100%
	Tenancy household data	96%
	Register data	99%
Inland Revenue	EMS	100%
	IR4	100%
Ministry of Education	Student enrolment	97%
	Student interventions	97%
	Student leavers	98%
	Student qualification	98%
	Completion (tertiary)	94%

[4] Black, A (2016). The IDI prototype spine's creation and coverage. (Statistics New Zealand Working Paper No 16–03). Retrieved from www.stats.govt.nz.

Data set source	Data set name	Spine match rate
	Enrolment (tertiary)	93%
	Industry training	97%
Ministry of Health	Hospitalisations	98%
	MH service activity (PRIMHD)	99%
	Pharmaceuticals	96%
Ministry of Social Development	Spells	100%
	Partner	100%
	Incapacity	100%
	First Tier Expenditure	100%
	Second Tier Expenditure	100%
	Third Tier Expenditure	100%
Police	Police proceedings	95%
Working for Families	Spells (WFF)	100%

5.3.2 Reliability

Standard investigations that we performed on the quality of data include:

- Checks on internal consistency of rate files
- Consistency across files that contain similar information
- Consistency with files used in previous years, or else consistency with publicly available data
- Proportion of people linking to the spine
- Reviewing trends through time.

All the data that has been used in this exercise is similar to the data that have been used since the 2020 round of modelling.

The population data were compared with Statistics NZ's estimated resident population, with differences investigated to understand any reasons why. Differences are due to the slightly different approach and were within the bounds of our expectations. Additionally, the population numbers were compared with Treasury population numbers. Our approach for defining the resident population is based on the Treasury approach (see Section 5.6.1 for details). The numbers reconciled with those of Treasury.

Income data was compared against publicly available data released by Inland Revenue. No material differences were found. The income data used are sourced from Statistics NZ derived data, which, based on discussions with Statistics NZ, were checked extensively by them. We have also confirmed that the derived data are consistent with the Inland Revenue tables inside the IDI.

WFF has not been updated since the 2020 model due to changes in the data making it unreliable. Discussions with Stats NZ have suggested that this is a result of changes to Inland Revenue systems as part of their Business Transformation Programme. There was no available solution to the data issue at the time of the modelling. WFF is not an input into the model and is only used as a modelling output, so we have retained the WFF model from previous modelling rounds.

For the 2020 modelling round we compared industry data against two data sources released by Stats NZ. These were:

- The Household Labour Force Survey (HLFS) people employed by industry.
- Employment indicator series which is based on EMS and payday filing data.

The two Stats NZ datasets have different counts for some industries due to the different sources used, and in particular, the differing inclusion of people who are not in the EMS. However, in each case our prepared data inside the IDI is materially close enough to be reliable for our purposes. The industry data within the IDI has not materially changed across refreshes since the 2020 modelling round.

Furthermore, for data that was used in the previous model we have compared key data sources against the 2023 data. This has allowed for a secondary data quality assurance check, by testing the robustness of our data preparation code to changes in the tables in the IDI between data refreshes. A summary of the percentage differences between refreshes for key prepared variables is provided in Table 5.3. The table shows only a subset of the variables that were compared. Extensive comparisons were performed for all prepared variables across different periods rather than just the aggregate information shown in the table below.

Table 5.3 – Summary of comparison between October 2022 and June 2023 data refreshes

Variable	Value	2023 refresh /2022 refresh
Housing State	Public Housing	100%
Housing State	Accommodation supplement	100%
Housing State	NIL	100%
Register flag		99%
Benefit state	A main benefit	100%
Benefit state	Supplementary only	100%
Benefit state	Not on benefit	100%
Care and Protection	Any interaction	100%
Youth Justice	Any interaction	100%
Sex	Male	98%
Sex	Female	98%
Income band	1	93%
Income band	2	100%
Income band	3	99%
Income band	4	101%
Deaths		100%
Tertiary enrolment		100%
Mental health flag		101%

Variable	Value	2023 refresh /2022 refresh
Committing an offence		100%
Corrections activity in past 10 years		99%
Population		98%

Note that while we make significant efforts to check the quality of data used in our analysis, we do not take ultimate responsibility for the accuracy and completeness of the data. Our reliance on the data provided is further discussed in Section 14.

The population decreases in the order of 1-2% seen above are the result of a removal of people from the spine who had been incorrectly added in the refresh used in the 2022 modelling round. This resulted in around 100k people being removed from our resident population, returning it closer to 2021 levels. The change did not necessitate any specific modelling changes but may have had some impact on modelling results.

Based on our checks and reviews we believe the datasets are sufficiently accurate, consistent and coherent; and we are satisfied that they are appropriate for use in this project.

5.3.3 Missing values

In a few cases, missing values were encountered in the data for key demographic variables such as sex, ethnicity, date of birth and location. For the missing values an imputation method was performed using stratified sampling to ensure that this information was not missing for anyone.

As well as for demographic information the following MSD-related fields contained missing values:

- Incapacity (type and number)
- SLP reassessment frequency

In some projection models, missing values are reasonable and can be included in the modelling process as an extra categorical level. For the benefit transition models, one of the main causes of missing incapacity or reassessment details is a fast exit from the benefit system (suggesting perhaps that there was insufficient time in these instances to collect client information fully). This means that missing values appear to predict a fast exit from the benefit system, when in fact the reverse is true (fast exits lead to missing values).

To avoid this bias, we have interpolated these missing MSD data values; that is, we randomly allocated values in cases where they were missing. This allocation was performed based on the distribution of variables for the clients with non-missing values when they first enter the benefit system. We believe this is the most effective way of handling missing values and avoids the need to delete these records entirely when fitting models. Extra check variables were created to indicate when variables had been interpolated.

There were a number of other variables across other domains that had missing values. These other variables were not imputed and were instead left as missing.

5.3.4 Public housing data

As for previous years, data on the public housing system have materially lower quality than those on the benefit system. More detail is included in previous reports with most of this still applicable inside the IDI.

The move to the IDI resulted in further issues with the public housing data. The following table shows the major issues with the data inside the IDI and what was done to remediate the issues. Our approach to these issues has not changed since the 2019 model.

Table 5.4 – List of issues with the public housing data in the IDI

Data set source	Description
There are approximately 30,000 children missing from the June 2019 IDI refresh.	In earlier IDI refreshes these children were incorrectly given the same unique identifier, these children have now been removed from the data. We added these children back into the dataset using data from an earlier refresh however we are missing data for more recent quarters and the linking of individuals across IDI refreshes is not a perfect process. We are not projecting children in housing in the model, so we do not expect this issue to have a material impact on the simulation model.
Poor linking across the data migration that occurred from Kāinga Ora to MSD in September 2015. This mainly affected children.	We have attempted to link these individuals across the data migration dates using demographic information such as date of birth.
Significant drop-off in number of people in housing from the Kāinga Ora data to the MSD data (i.e. in September 2015).	This is due to the lack of people in “evidence items” included. From discussions with MSD they believe that the evidence items do not give an accurate picture of who is in housing.
No house information is included for community housing provider houses	We know the location of the house through the tenancy data. For houses that transferred to community housing providers from Kāinga Ora we have been able to link these together.
Large decrease in the number of people in public housing from April 2009 to October 2010.	Separate data for this period was included by Statistics NZ in the sandpit. The linking rates are low for this data, so we have manually linked most of these people using demographic and household information.
There are some primary tenants who are in two houses at the same time.	The tenant is kept in the first house they moved in to.
Some individuals in public housing get removed temporarily when they reach the age of 16.	This appears to be when an individual with a child SWN is allocated an adult SWN and they go missing. We have patched up where we can.

We have liaised heavily with Statistics NZ and MSD about these issues.

5.3.5 Other data issues

There were additional issues with some datasets for the 2023 modelling round. These were:

- **Corrections data** – the corrections data the in June 2023 refresh has some errors affecting data over the years 2013-2020 which made it unusable for modelling. To fix this we used the corrections data from the October 2022 refresh (the last correct refresh) as far as it went (to 30 June 2022). We added on the data from the June 2023 refresh for time after this, as this period of the data is unaffected by the data quality issue. Linking between refreshes means the correction spells in the 2023 modelling round was slightly lower than in 2022 modelling round despite using the same dataset, however this difference was deemed to be inconsequential for modelling.

We spoke with people at corrections who informed us this will be corrected in future refreshes.

- **PRIMHD data** – for the 2022 model we identified that the PRIMHD data looked low for the most recent 12 months available, up to 30 June 2022. The dataset has not been updated since then. We understand this is likely due to underreporting by certain service providers or other data quality issues. We therefore ignored the last year of PRIMHD data and imputed from 1 July 2021 onwards.

5.3.6 Roll forward of datasets for modelling

In the IDI, data are gathered from several different sources. The frequency of data supply to Statistics New Zealand and data extract dates vary across data providers. Data from some sources end before the welfare and housing data (see section 5.2). For these data sources, we extend the data series to 31 March 2023 using predictive modelling to sample pathways that are consistent with actual data to 31 March 2023. The datasets that have been adjusted in this way were:

- **Secondary education** – all secondary school leavers information was rolled forward. Additionally, it was imputed for all people up to age 30 (inclusive) who didn't have secondary school information. These were performed by sampling from historical data for people with similar characteristics.
- **Tertiary education** – all tertiary education information was rolled forward for one quarter. This was performed by sampling from historical data for people with similar characteristics. Data on completions needed to be rolled forward a further year and was done by using a GBM to assign completions to people enrolled in tertiary education.
- **Income** – annual tax return information (sole trader, partnership, company) for the year to March 2023 was rolled forward. This was done by creating two separate Gradient Boosted Machine models (GBMs) for each time period. First, to estimate whether someone would have had additional annual tax return information. Secondly, to estimate what the amount was if they indeed had some. These amounts were then spread evenly across any quarters in the tax year.
- **Industry** – current industry was imputed for those who had been assigned some income for the year to March 2023 (i.e. income roll forward step as above) but did not have an income recorded in the EMS data, so had no industry recorded. This was done by using a GBM to assign an industry to individuals based on various characteristics such as income history, previous industry and age. Other industry variables (e.g. industry duration and last industry) were rolled forward using the imputed industry.
- **Hospitalisations** – acute hospitalisations were rolled forward using GBMs fit to historical data based on rolling forward one quarter at a time.
- **Mental health** –we rolled forward pharmaceutical data for one quarter and PRIMHD data for seven quarters. We categorised people based on their mental health history (e.g. past service events) and information we do know up to the modelling date (e.g. benefit status) and sampled from people with similar characteristics from prior data.

5.3.7 Roll forward of datasets for the starting population

Stats NZ have introduced regular refreshes to the IDI of benefit system data and the Inland Revenue EMS table. The regular refreshes have enabled us to run the simulation model as at 30 September 2023 instead of the refresh date of 31 March 2023. However, we have not used this data for fitting models.

For data outside the benefit system and inland revenue we took a variety of different approaches to bring the data forward. These were:

- **Population –** we have taken the population as of 31 March 2023, aged everyone up, and added on those who were receiving a benefit in the June and September 2023 quarters.

- **Public housing –** We have assumed everyone in public housing on 31 March 2023 remains in public housing unless they received the accommodation supplement. This places limitations on the ability to analyse recent public housing exits through the modelling.
- **Income –** We have applied the same approach as for modelling for the annual tax return income sources. For those who could not be matched from the monthly income data to the monthly benefit data we have rolled forward their income variables using our fitted models.
- **Education and CYF-** We rolled forward education and CYF data using our fitted models. This was important as it is heavily dependent on age.
- **Corrections and District -**We rolled forward corrections and district data using our fitted models.
- **Industry –** For people who earned income in the March 2023 quarter we used their same industry. For those who were not earning income we used their last industry. For everyone left over we used a sampling approach stratified on age, ethnicity, sex, district and earned income amount.
- **Other sources -** For all other sources we used the data as at 31 March 2023. Where people could not be linked to the June refresh data we have sampled their remaining variables based on demographic and/or benefit state and income band where available.

5.4 Privacy

All data analysis for this report has been produced in accordance with New Zealand legislation and frameworks guiding official statistics. The production of official statistics in New Zealand is guided by:

- The Statistics Act 1975
- Other legislation, such as the:
  - Privacy Act 1993
  - Official Information Act 1982
  - Public Records Act 2005
- New Zealand Data and Information Management Principles

We have also ensured we are in keeping with the following policies and protocols from Statistics NZ:

- Privacy and confidentiality guidelines
- Data integration guidelines
- Information privacy, security and confidentiality policy.

5.5 Child entrants from the Children's model

Child entrants are the main source of future entrants to the population and are sourced from the 2022 children's model. The children's model itself is a micro-simulation of the child population in New Zealand, so is available at the granular level of one row per child turning 16 in each of forty quarters from December 2023 to June 2032.

Individual characteristics are available for these children including sex, ethnicity, education, any public housing or benefits history including inter-generational welfare etc. Some significant adjustments were needed to:

- Ensure the data contains the same variables as the main projection cohort
- Ensure the distribution of key variables is sensible against current 16-year-olds.

Variables from the children's model that needed adjusting to be in the same form as our projection cohort include:

- Location: this is not projected so we have assumed the same location as when projection started.
- Intergenerational variables: we calculate intergenerational variables based on whether they have a parent supported by a benefit in each quarter from age 13-18. Our approach is to look forward in the children's model to when the child turns 18 rather than stopping at age 16 when they enter the model.
- When they left school and their highest NZQF level on leaving school. As above we look forward in the projections to when they leave school rather than capping at age 16.
- Other variables such as benefit-related variables and housing-related variables. These are imputed by sampling from the historical data for people with similar characteristics.

Variables that we needed to correct distributions for were adjusted by stratified sampling from characteristics of actual 16-year-olds:

- Ethnicity was imputed for those with missing ethnicity. Sampling was performed by sex, CYF type, whether their parents had been supported by a benefit from age 13-18 and whether they had been suspended or stood down.
- Education-related variables including age exiting education and NZQF level at exit. This was done because the NZQF level on finishing school in the 2022 children's model did not fit the distribution in the historical data. We have discussed this issue with OT previously. The education variables were sampled by sex, ethnicity, CYF type, whether parents were supported by a benefit when aged 13-18 and whether they had been suspended/stood down.

5.6 Data Preparation

The source data used in the model required extensive transformation to convert them into the form required for building and running the model. The data preparation process transforms the data from the source data into a suitable form.

Aspects that are consistent across each of the source data domains include:

- Producing an aggregate dataset from variables across the source datasets in that domain by linking on each person.
- In most cases then aggregating the dataset to be one row per person per quarter. Some datasets (e.g. demographics) are just one row per person.
- Transforming the values in the source data into the variables chosen for modelling. This sometimes includes aggregating over history.
- Performing data quality checking and cleansing throughout the process.

5.6.1 Population and modelling spine

Our approach to defining the resident population within the IDI is based on the approach taken by Treasury[5]. The Treasury approach builds on the methodology developed by Statistics New Zealand[6]. The approach involves taking all people who:

---

[5] McLeod, K (2018). *Where We Come from, Where We Go – Describing Population Change in New Zealand (AP 18/02)*. Retrieved from https://treasury.govt.nz/publications/ap/ap-18-02-html

[6] Gibb, S, Bycroft, C, Matheson-Dunning, N (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI). Retrieved from https://www.stats.govt.nz/assets/Research/Identifying-the-New-Zealand-resident-population-in-the-Integrated-Data-Infrastructure/identifying-nz-resident-population-in-idi.pdf

- are on the IDI spine (collection of birth records, IRD numbers and student, work and resident visa applications),
- have a sign of activity (e.g. education enrolment, utilisation of health system, paid tax) since 2008,
- are alive, and
- have spent sufficient time in the country.

Both methodologies give population estimates which are similar to official population statistics.

The main difference between the Treasury and Statistics New Zealand methodologies is that:

- Treasury check for any sign of activity after 30 June 2008 and requires a person to be in the country for 12 out of 16 months centred around the reference date
- Statistics New Zealand check for any sign of activity in the year of interest and requires a person to be in the country for at least 2 out of the 12 months centred around the reference date.

Under the Treasury approach, a person can only leave the population due to death or migration. It is possible for people to leave the population due to inactivity in the Statistics New Zealand approach. As such the population under the Treasury approach is slightly more stable.

It is important to note that the resident population cannot be considered reliable prior to 2008 because there are fewer data sources available inside the IDI to give a sign of activity.

As well as the population dataset we also have a modelling spine. The difference between the two is that the modelling spine includes all people in the population as well as everyone who is in public housing or receiving a benefit in the quarter and is not in the population. This is approximately an extra 20,000 people per quarter. The main reason why these 20,000 people are not already estimated to be in the population is because they have not yet satisfied the ‘at least 12 of the last 16 months in NZ’ part of the estimated resident population definition.

The modelling spine has been used as the basis for our modelling datasets and our starting dataset for feeding into the projection.

5.6.2 Demographics

The main source of demographic information that has been used is the personal details table which is derived by Statistics NZ inside the IDI. This is used as it is derived using the most reliable sources inside the various tables inside the IDI. Variables from this table used are:

- Sex
- Birth year and birth month
- Deceased year and deceased month
- Ethnicity (discussed further below)

Kāinga Ora tables are not used as a source for the personal details table. Hence, in some cases where someone is in housing their demographic information is not in the personal details table. For these individuals we have used their demographic information directly from the housing tables. This includes all of the above information excluding a record of when someone is deceased.

5.6.2.1 Ethnicity

Ethnicity is recorded in a number of different data sources inside the IDI by a number of different agencies. Sometimes these sources will change how ethnicity is recorded over time. We have used the ‘source ranked ethnicity’ from the personal details table derived by Statistics NZ, which places greater weight on sources that have been shown to be similar to census information (i.e. self-identified ethnicity). The approach that Statistics NZ have taken determines whether an individual has ever identified as Māori, Pacific Island, NZ/European, Asian, Middle Eastern/Latin American/African and/or other.

Consistent with previous years, we have used priority ranked ethnicity in our models with the ethnic group determined by:

- If identified as Māori, then Māori
- Else, if Pacific Island then Pacific Island
- Else, if Asian then Asian
- Else, if Other or Middle Eastern/Latin American/African then Other
- Else, NZ/European.

We note that it is now commonplace to reference ethnicity on a total ethnicity basis rather than a priority ranked ethnicity basis. While the models use priority ranked ethnicity as a predictor, we report modelling results on a total ethnicity basis.

5.6.3 Addresses

Similar to the personal details table for demographics, Stats NZ has derived an 'address notification' table which contains its best estimate of an individual's most likely address based on prioritisation of address sources. We have used this table for addresses with a few supplements:

- The table is missing some of the MSD addresses, mainly for older addresses. The address notification table has been supplemented with these addresses.
- A number of individuals in MSD's child dataset are MSD orphans who do not have any address data. The address of the caregiver is used for these cases.
- The address is overwritten by the address of the house for those who are in public housing.

Ultimately, the meshblock where someone is living at the end of the quarter is the key variable derived from the above approach. District group and the TLA code are then derived from the meshblock.

5.6.4 Benefits

Benefits data has been prepared consistently with previous valuations. Significant amounts of data are combined to produce the benefits table. A high-level summary of the process is as follows:

1. A spell file is created with periods that a person is main beneficiary, partner supported by benefit or child supported by benefit.
2. A rate period file is then created, which is from the spell file combined with the first tier and second tier expenditure information. This dataset contains payment amounts for each spell for each individual.
3. The rate period file is then put in a quarterly format. The benefit type someone received in the quarter is then determined according to the following rules in order:
   a. The main benefit that was received for the greatest number of days in the quarter. Where the days are tied the order applied is SLH, JHD, SPS, JWR, SLC then EMB.
   b. Orphan's benefit (ORP)
   c. Supplementary benefits (SUP)
   d. Otherwise, deemed 'not on benefit' (NOB).
4. From the rate period file aggregate the payment amount for each benefit type per person per quarter.
5. Combine child dataset with benefit definition derived in step 3 to calculate the proportion of time between age 13 and 18 where they had a parent supported by a main benefit (intergenerational variables).
6. Create an MSD demographic table. This is a result of combining the following from different datasets inside the IDI:
   a. Partner information
   b. Child information

   c. MSD district information
   d. Incapacity information
   e. Education information held by MSD (highest qualification from SWN file)
   f. SLP reassessment frequency.

   In this step missing values for the last four items are forward/backfilled and imputed where still missing.

7. History benefit variables are calculated by looking back through each individual's history supported by benefit.
8. All of the above tables are then combined to produce a benefit dataset.

In 2013, welfare reforms were made which resulted in different benefit types and amounts. We modify data prior to the welfare reforms to reflect the post-reform world. This is to ensure all our welfare data is on a similar basis.

5.6.4.1 Key variables

- Benefit in the quarter: what benefit is received in the quarter and whether they are supported by a benefit at the end of the quarter. Determined according to the rules in step 3 above.
- Benefit history: Number of quarters on current benefit, previous benefit, quarters since first benefit, proportion of time supported by a main benefit in past 3 and 5 years and spent in each benefit state.
- Client intergenerational history: Whether the client's parents were beneficiaries while the client was aged 13-18 and the intensity of benefit receipt. Defined up to the age of 30 (inclusive).
- Family-related variables: Youngest child age and number of registered children (for all main benefit clients), and partner flag (for JS and SLP clients).
- Health and disability-related variables: Incapacity type for JS-HCD and SLP-HCD clients, HCD reassessment frequency and whether the incapacity belongs to the primary client or to their partner.
- Payments: The total amount received as a benefit in the quarter split by the different types of benefit.

5.6.5 Education

Education is split into secondary and tertiary with the dataset sources and data preparation being different for both.

The output table from the secondary school data has one row per person. It is created from the student leavers, student qualification, and student intervention datasets.

The tertiary education table contains one row per person per quarter, consistent with most other datasets. A record is obtained from the 'enrolments' dataset of the highest qualification that the individual has attempted as at the end of the quarter. A record is obtained from the 'completions' dataset of the highest qualification that the individual has attained as at the end of the quarter. This is done for both tertiary education and industry training and then combined as our tertiary education table. Where we refer to tertiary education in this report it refers to tertiary education and industry training combined.

The variables derived in the education data are:

- Whether a person has left school.
- The highest NZQF level attained when leaving school (not retained after age 30).
- The total duration of any stand-downs or suspensions while at school.
- Whether the person is enrolled in tertiary education.
- The duration of the current enrolment status – this is either the time they have been enrolled in tertiary education consecutively, or the time they were last enrolled.
- The highest NZQF level of any tertiary qualification enrolled (not retained after age 30).

- The highest NZQF level of any tertiary qualification attained (not retained after age 30).

5.6.6 Oranga Tamariki data

Oranga Tamariki (OT) data has been prepared in the same manner as the previous report. The variable definitions align with those used in the Oranga Tamariki's children model.

OT events have been combined from the different OT tables inside the IDI. These events include:

- Reports of concern with no further action (intakes table)
- Investigations
- Family group conferences
- Youth high court
- Placements.

These different events are then aggregated to produce the following variables:

- CNP status - the highest level of intervention for child protection
- YJU status - the highest level of intervention for Youth Justice
- Number of CNP days – the total number of days in child protection placements in the quarter
- Number of YJU days – the total number of days in Youth Justice placements in the quarter.

These variables are not retained after age 30.

5.6.7 Income

Personal income data used in the modelling is directly from Statistics NZ's derived income table by tax year. This table is created from several different Inland Revenue tables inside the IDI. These are:

- Monthly records from the Employer Monthly Schedule (EMS)
- Self-employed income from the IR3
- Rental income from the IR3
- Company director/shareholder income from the IR4S
- Partnership income from the IR20.

Income information from the EMS includes wages and salaries, withholding payments, ACC weekly compensation, Student Allowance, paid parental leave, welfare benefits and NZ superannuation.

Income information that is not from the EMS is recorded only on a tax (March) year basis. These income amounts have been split quarterly in proportion to the time the individual has spent not supported by a main benefit over the year. This has added some additional seasonality to the income amounts with small income steps in June. Seasonality is included as a factor in income models, so we have reflected this seasonality in the projections. Overall, it results in income from quarter to quarter that might not match the pattern of how it has been earned.

The main income modelling variable includes all of the above except NZ Superannuation and welfare benefits. NZ Superannuation is included in a separate variable.

The main income amount has been banded into 4 levels for modelling purposes. More information is included in Section 11.1.

As well as income band and income amounts, we keep track of the following variables:

- The number of quarters an individual has been in their current band
- The number of quarters since an individual has been in income band 1 (i.e. had zero income)

- The income band an individual was in prior to their current band
- Total income earned in the past year
- Total income earned in the past three years.

WFF tax credits have not been updated since the 2020 model due to data issues explained in Section 5.3.2. In previous years, they were derived from the family tax return details table inside the IDI. Assumptions made include the following:

- The amount has been allocated to the primary individual on the return. In the case of a couple this means only one person is allocated the tax credit.
- Return information is on a March year basis. We have split this into quarterly amounts by dividing by four. This is a slightly different approach to that noted above for spreading annual income amounts (i.e. based on time spent not supported by benefit), as people are entitled to some tax credits while being supported by a benefit.

Income has been inflated into current values using average weekly earnings (AWE).

5.6.8 Industry

There are 17 different industry groupings used in the model. These are:

- Forestry and mining (combining 'Agriculture, forestry and fishing' and 'Mining' ANZSIC06 codes)
- Manufacturing
- Electricity, gas, water and waste services
- Construction
- Wholesale trade
- Retail trade
- Accommodation and food services
- Transport, postal and warehousing
- Information media and telecommunications
- Financial and insurance services
- Rental, hiring and real estate services
- Professional, scientific, technical, administrative and support services (combining 'Professional, scientific and technical services' and 'Administrative and support services' ANZSIC06 codes)
- Public administration and safety
- Education and training
- Health care and social assistance
- Arts, recreation and other services (combining 'Arts and recreation services' and 'Other services' ANZSIC06 codes)
- Other.

The first sixteen categories are consistent with what Stats NZ publishes for their employment and earnings reports.

The 'other' category represents people who are earning income but there is no industry record within the IDI. These people are all receiving only the following:

- Student allowance
- Paid parental leave
- Weekly compensation from ACC
- Partnership/company income but no record in steps 2-4 below.

Industry is derived from various tables within the IDI and applied in a priority order. This order is:

1. Industry from monthly records from the Employer Monthly Schedule (EMS).
2. Employer records from the Stats NZ derived tables and then an employer-industry mapping from the EMS table. The mapping uses the industry for the most employees for that employer in the quarter.
3. Employer records from the Stats NZ derived tables and then matching the employer to the business register and the industry from the business register tables. There are very few cases with multiple records in the business register for the same employer, in these cases we randomly allocate the business register record.
4. Industry record from the Inland Revenue Customers table which is gathered from ACC levy records.

Where someone has worked for multiple employers in the same quarter, we take the industry from the employer they earned the most income from the in quarter. The ‘Other’ industry group is only used as a last resort (i.e. it is only used if there are no other industry records for that person in the quarter).

Industry is recorded in the EMS table on both an enterprise and PBN (permanent business number) level. An employer has only one enterprise number but can have multiple different PBNs (for example, for different sites). These different PBNs can then have different industries. We have used industry from the PBN as it more takes into account the individual circumstances of where the employee is based.

Two other variables are derived from the industry:

- The last industry someone worked in is defined as the industry they were in when they last earned income
- The duration (number of quarters) someone has worked in that industry consecutively. This resets to zero if the individual either stops earning income for a full quarter or changes industry.

5.6.9 Housing

Significant amounts of cleaning were needed for the housing data before they were ready for use in data preparation. More detail on this is included in Section 5.3.4.

Housing data preparation is consistent with what has been done in previous years. The following steps outline the process:

1. Each household/application/house is given a consistent ID. The raw data contain up to three different IDs, depending on what system the data originally came from.
2. Household data are prepared on a one row per household per quarter basis containing the Household ID, House ID and variables that are the same for the household such as market rents and subsidies.
3. Individual data are prepared on a one row per person per quarter basis containing information on the individual, the Household ID and the relationship to the primary tenant. A significant amount of cleaning goes on in this step, including:

    a. Combining IDs where someone has the same demographic information and is on the spine. This is particularly important across data migration periods.
    b. Back-filling and forward-filling information such as relationship to the primary tenant when it is missing.
    c. Working out the primary tenant when there is either zero or more than one primary tenant recorded in the house.

4. The household (from step 2), the individual data (from step 3) and the houses data are then combined to get a complete record of one row per person per quarter.
5. Using a similar approach to steps 2-3, the register data is prepared and combined. A further step is needed here to match the application preferences with suburbs in New Zealand to get these at a meshblock level.
6. The register and housing data are then combined, and various historical variables are then calculated for use in the model.
7. The housing state (h_state) is then derived from the above data and through linking to the benefit dataset. This is defined as:

a. If in public housing at all in the quarter, then “PH”; else
b. If received any Accommodation Supplement in the quarter, then “AS”; else
c. “NIL”

Key variables produced from the housing data preparation are:

- Public housing variables: Whether a client is in public housing, whether a primary tenant or a partner, household size and other household characteristics.
- Amount variables: The amount of the subsidy, market rent and income. All amounts are inflated at benefit inflation rates.
- Register related variables: Whether a client is part of an application on the register, whether they are the primary applicant, criteria scores for priority in public housing.
- Historical variables: Time in public housing, time in current house, time since being on the register, time since being in public housing, historical variables related to previous spells in public housing.

5.6.10 Justice Sector

Justice data is made up of corrections and offences leading to police proceedings.

Corrections periods are obtained from the Major MGMT Periods dataset inside the IDI. In this dataset, individual periods for an offender are rolled up into a single chronological view. If periods occur concurrently, trumping rules apply so that the most serious sentences are prioritised.

Corrections spells are grouped into the following:

- Prison (in prison or remand)
- Offences excluding driving
- Theft offences.

The total number of days in a quarter that an individual has been in corrections for any of the above is then recorded in separate variables. These are then used to derive the corrections variables that are used in the model:

- Percentage of time in prison in the previous year.
- Percentage of time in the previous year serving any criminal sentence, excluding sentences for driving-related offences.
- Of the past 10 years, percentage of time serving any type of criminal sentence, excluding sentences for driving-related offences.
- Of the past 10 years, percentage of time serving a criminal sentence relating to a theft offence.

Number and type of alleged offences are obtained from the police proceedings post-count data. This dataset counts a person once on each day they are proceeded against by the police. This includes court and non-court action. Where there are multiple offences in a day the principal offence is determined by the relative rankings of the offence seriousness. This approach is consistent with the Justice Sector Investment Approach model.

The number of offences and the duration in their current offence state are then aggregated by quarter. If an individual committed an offence in the quarter, then the duration represents the number of consecutive quarters the individual has committed at least one offence. If they did not commit an offence in the quarter, then it is the number of quarters since the last offence.

A separate dataset is produced which lists the offence type by person, by quarter, by offence. Offence types are grouped into seven categories including some split by high/low seriousness. Refer to Section 12.3.2 for more detail.

5.6.11 Health

Health data is aggregated from numerous sources from within the IDI.

We previously relied on the Social Wellbeing Agency's (SWA) mental health data definition to define mental health and addiction service events for the New Zealand resident population. This definition captures mental health conditions as well as addiction service access or treatment. The diagnosis categories provided by the definition are not exhaustive and should be treated as simple high-level inferences made from the IDI data. Since the 2020 model we have used an adjusted definition based on discussions with the Ministry of Health. For details of the changes, refer to the 2020 Technical Report. Our current definition utilises data from the following sources:

- PRIMHD – The Ministry of Health collection of national mental health conditions and addiction information on service activity and outcomes for healthcare users. The data from this source is separated into inpatient and community service events.
- Pharmaceutical dispensing data – Ministry of Health and PHARMAC data containing claims and payment information from pharmacists for subsidised medication.

The SWA definition also includes hospitalisation and MSD incapacity data, but we exclude these from our definition.

Variables that have been derived for modelling include:

- A flag if there is a mental health or addiction service event in the quarter.
- The number of consecutive quarters a person has had mental health or addiction service events or the number of quarters since their last event.
- Three flags for the type of activity for a mental health event. These three flags are for:
    1. Pharmaceuticals
    2. Community mental health specialist events (from PRIMHD)
    3. Inpatient mental health specialist events (from PRIMHD)
- A highest mental health event variable which is ranked in the order above.
- For each of the flags there is an equivalent duration variable that measures the number of consecutive quarters a person has had that type of event or the number of quarters since their last event of that type.
- Whether a person with an acute hospitalisation is discharged in the quarter, as well as the number of days spent in hospital (which may span more than one quarter).
- The number of consecutive quarters with acute hospital discharge events or non-events.

5.7 Summary of changes since the previous report

- Data used for modelling was prepared up to 31 March instead of 30 June as done for the 2022 modelling round.
- The resident population has reverted back to the 2021 level as those who were deleted from the IDI spine in 2022 have now been reverted.
- There were issues with corrections data this year which meant a combination of the October 2023 refresh and the June 2023 refresh was used to create the relevant corrections variables.
- We added imputation of industry for those who had been imputed to have some income for the year to March 2023 but did not have an income recorded in the EMS data (so had no industry recorded). In previous modelling rounds we did not impute these values.

- No imputation was required for the corrections/justice data for the 2022 modelling round, but was for the 2023 modelling round.

6 Assumptions

In this section we discuss the core assumptions relating to macroeconomic factors, the data and population formation, and some general assumptions that impact the modelling approach.

6.1 Economic assumptions

Various macroeconomic variables are used in the projection. These include:

- National and regional employment rates
- Future discount rates
- Future inflation rates (CPI and AWE)
- Market rent inflation forecasts.
- Net migration

Each of these assumptions is detailed below.

All inflation and rent assumptions are based on actual experience to 30 September 2023 and forecasts thereafter.

6.1.1 Unemployment rate forecasts

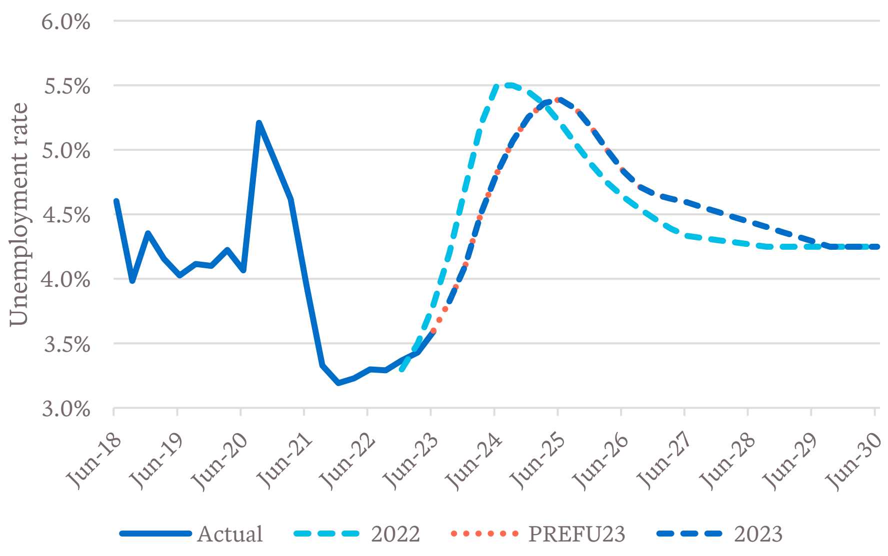
6.1.1.1 National rates

Figure 6.1 shows the adopted national unemployment rates used in the projection as at June 2023, together with actual rates and those assumed in the previous report for September 2022.

Treasury's Pre-election Economic and Fiscal Update (PREFU23) projects the unemployment rate to increase sharply until it reaches 5.4% in June 2025, before slowly decreasing towards our long-term assumption of 4.25%. We have adopted the Treasury forecast without modification.

The adopted rate in the mid-term is higher than that for the previous projection, reflecting a slower expected loosening of the currently tight labour market and a slower tapering of the tight monetary policy in response to the inflationary pressures.

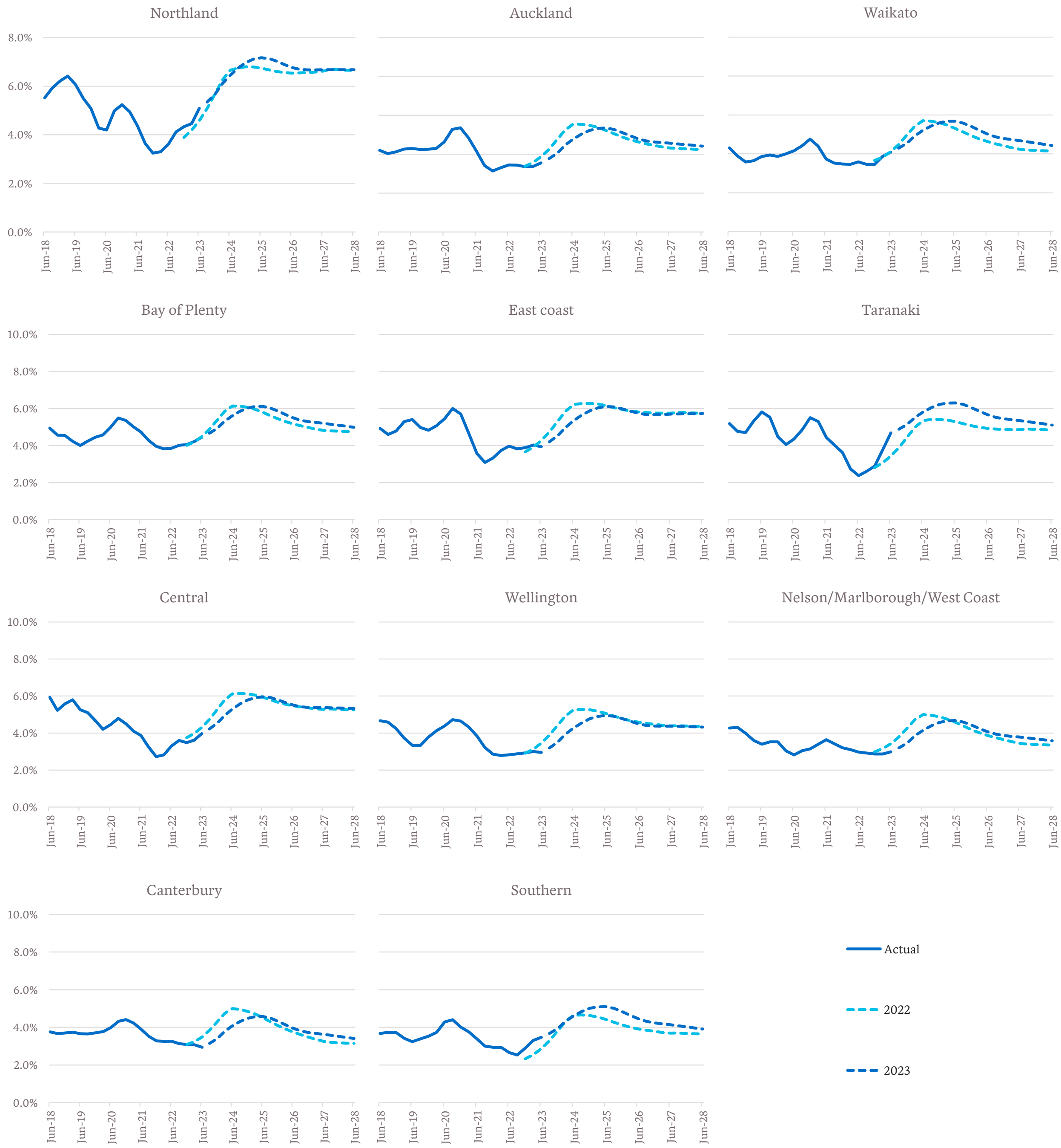
Figure 6.1 – National unemployment rate forecasts



6.1.1.2 Regional rates

Our models use regional rather than national unemployment rates for stronger predictive power. Treasury does not provide a forecast of regional rates so in previous years we developed a methodology to produce a forecast for these rates and have used the same methodology here. This is discussed in Appendix A. Regional rates broadly follow the shape of the national unemployment rate seen in Figure 6.1.

Figure 6.2 – Regional unemployment rates

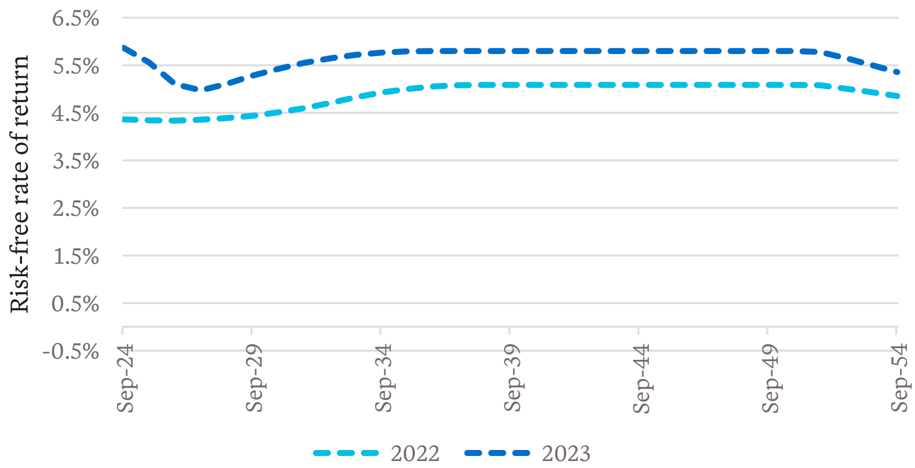


6.1.2 Discount rates

Discount rates are risk-free interest rates that reflect the time value of money. The projection uses risk free rates derived by Treasury using New Zealand Government bond rates. These are shown in Figure 6.3. Discount rates have increased over the period between the September 2022 and September 2023 modelling rounds.

A higher rate decreases the value of future cash flows but does not otherwise affect the projection.

Figure 6.3 – Discount rates forecast



6.1.3 Inflation rates

6.1.3.1 CPI

The rates of most benefit payments are now tied to AWE, which means AWE drives their level over time. In the medium to long-term, our AWE assumption is linked to CPI. Additionally, some supplementary benefit payments are tied to CPI. Thus, forecasts of future CPI levels are an important input to the projection of future payment levels. Recent CPI levels have been high so CPI inflation is now forecast to be higher in the short to medium term, reaching the long-term rate of 2% later.

Figure 6.4 – Short-term CPI inflation

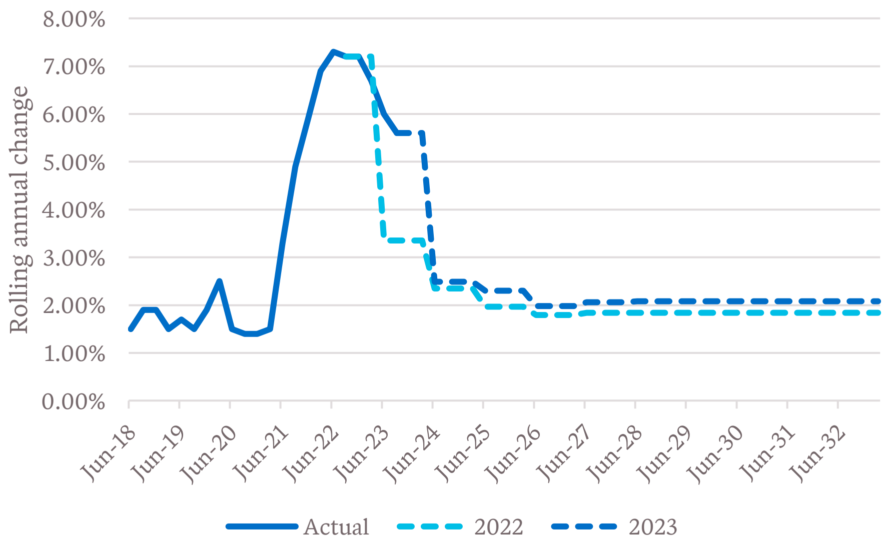
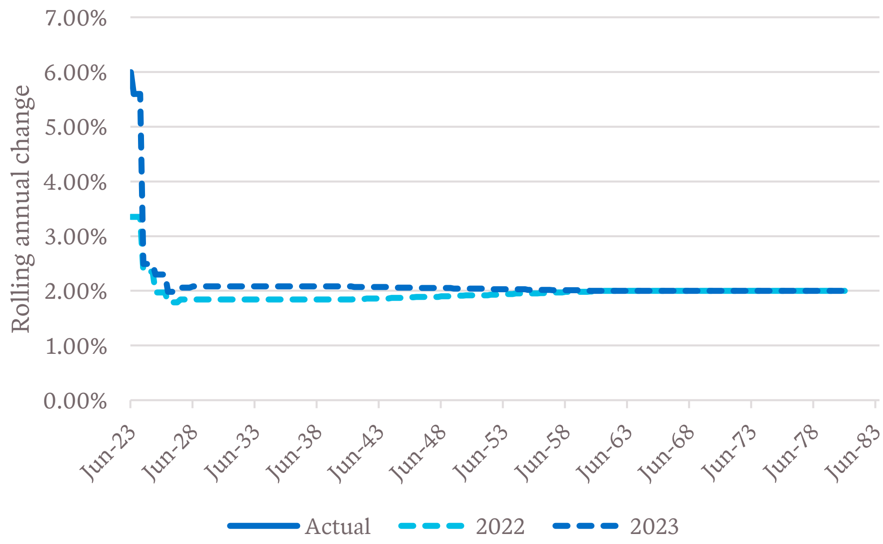


Figure 6.5 – Long-term CPI inflation



A higher inflation rate increases the value of future cash flows but does not otherwise affect the projection.

6.1.3.2 AWE

In addition to driving benefit payments rates, AWE also affects our rental growth assumption and pensioner incomes (which in turn affects IRRS in respect of pensioners). To forecast AWE, we use Treasury forecasts for as far into the future as they are available (up to 2027 in this case). Thereafter, we assume that AWE is forecast as CPI plus a margin (1.5% p.a. in the long term) and adjust the implied margin from the Treasury forecasts to the long-term margin over a six-year period.

AWE has maintained the strong levels seen since 2020. Low unemployment and high inflationary pressures over the last year have led to higher AWE forecasts in the short to medium term than those forecast for the 2022 model. In the long term, rates converge to the same rates. The forecasts, expressed as rolling annual rates, are shown in Figure 6.6 and Figure 6.7.

Figure 6.6 – Short-term AWE inflation

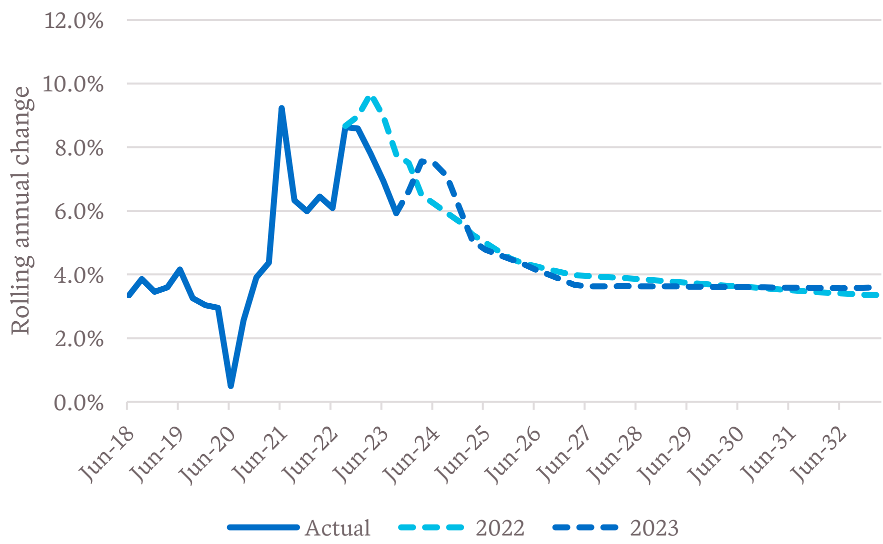
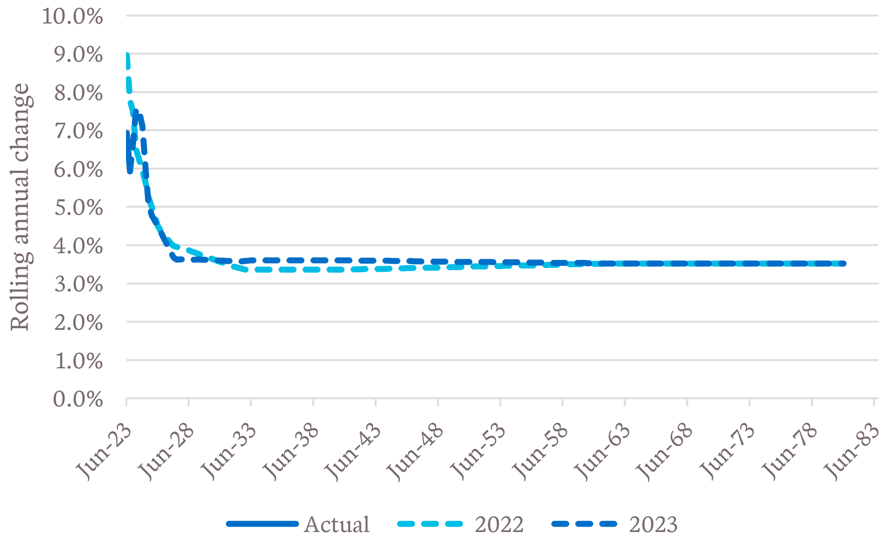


Figure 6.7 – Long-term AWE inflation



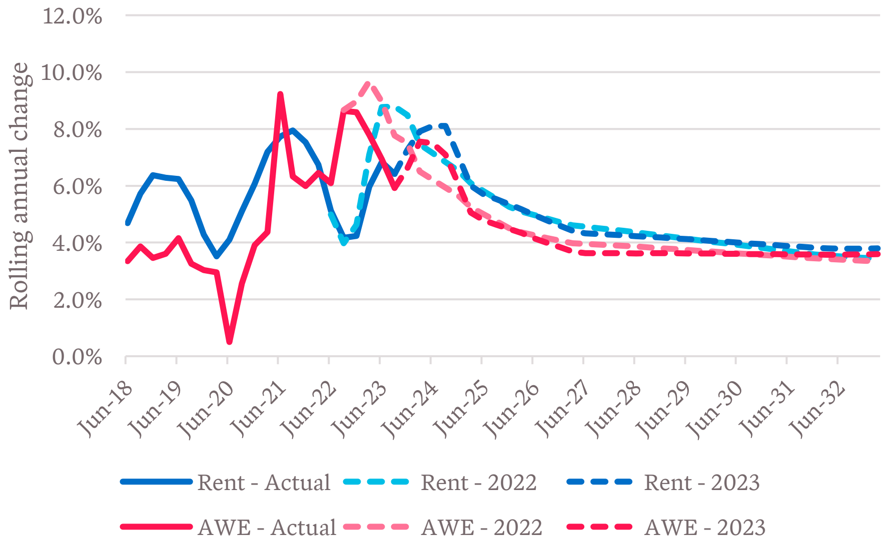
6.1.4 Market rent inflation

6.1.4.1 National rates

Market rent assumptions drive the level of IRRS over time and are very important to the public housing modelling results. Our models use first-quartile rents (these are closer to average social housing rents than the average or median), so we require assumptions for their inflation. We have assumed that growth in rents will be faster than AWE growth in the short to medium term as this has been the case in the recent past. There are a number of reasons why rents might grow faster than average wages. First, average wages may mask higher wage growth in some regions such as major cities. Second, housing costs can grow as a proportion of total income. Third, housing supply constraints can squeeze both the owner-occupier and rental markets higher. These supply constraints can be further compounded by population growth, both from births and migration.

Arguably there is less justification for assuming that growth in rents will be faster than AWE growth in the long term. Therefore, the growth rates in first quartile rents are set equal to AWE plus a premium for the first ten years (starting at +1%, linearly fading), so that after 10 years, rents grow at the same rate as AWE. Figure 6.8 shows these forecasts, with AWE also shown in the lower graph to provide context around their derivation.

Figure 6.8 – National rent inflation forecasts

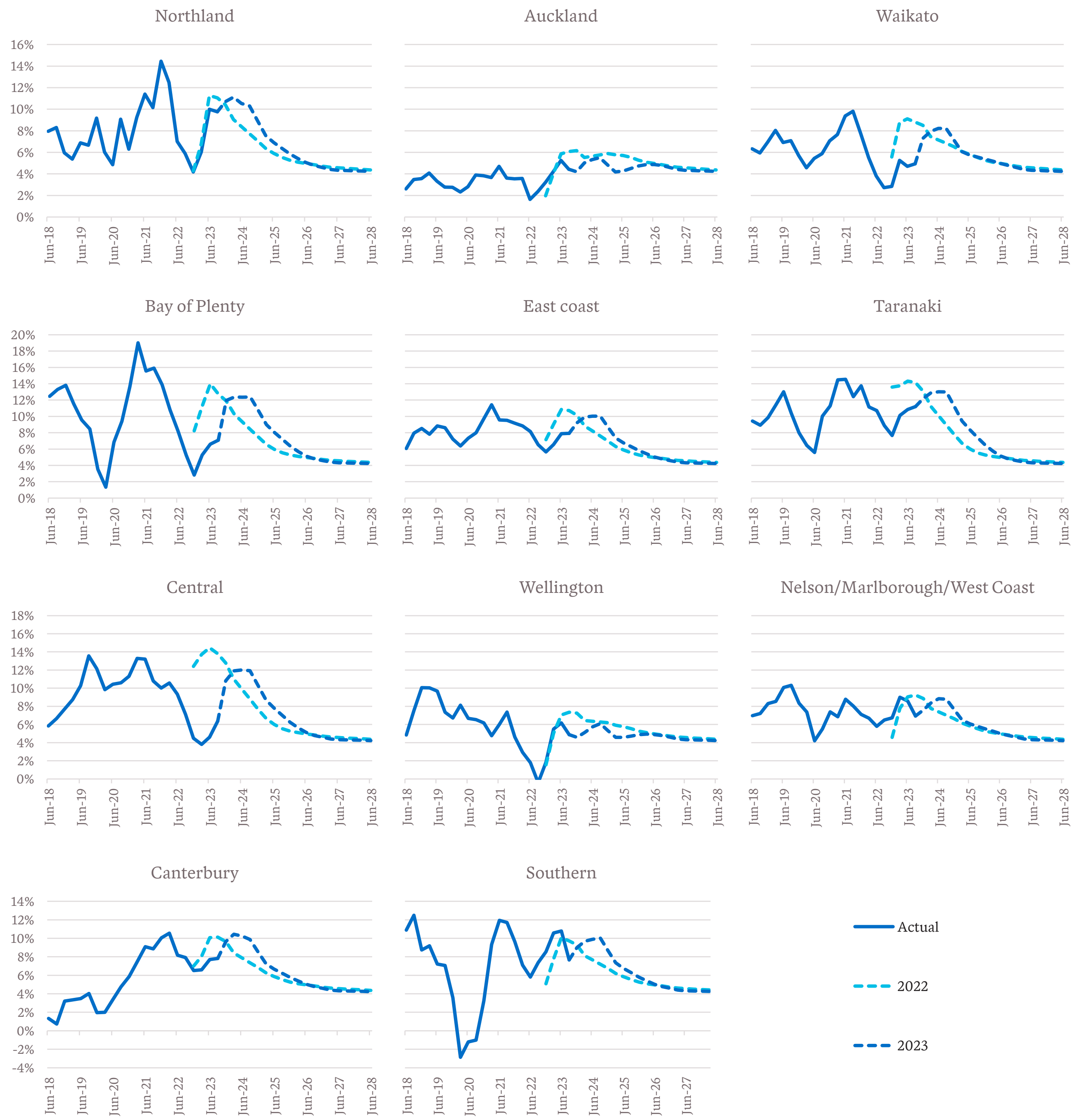


6.1.4.2 Regional rates

The housing models use regional rather than national rents, so forecasts of these rates are also required. Regional rates are assumed to have variations in the short-term reflecting regional experience but converge to the national growth rate after two years. The short-term variations are based upon the last two years of rent inflation in each region.

The national growth rate was lower than the 2022 forecast over the 12 months to September 2023. Rental growth has increased over the previous 12 months and is forecasted to increase again in the short to medium term. Figure 6.9 shows the observed rent inflation in all regions together with the current forecasts. The September 2022 forecasts are shown for comparison.

Figure 6.9 – Regional market rent inflation



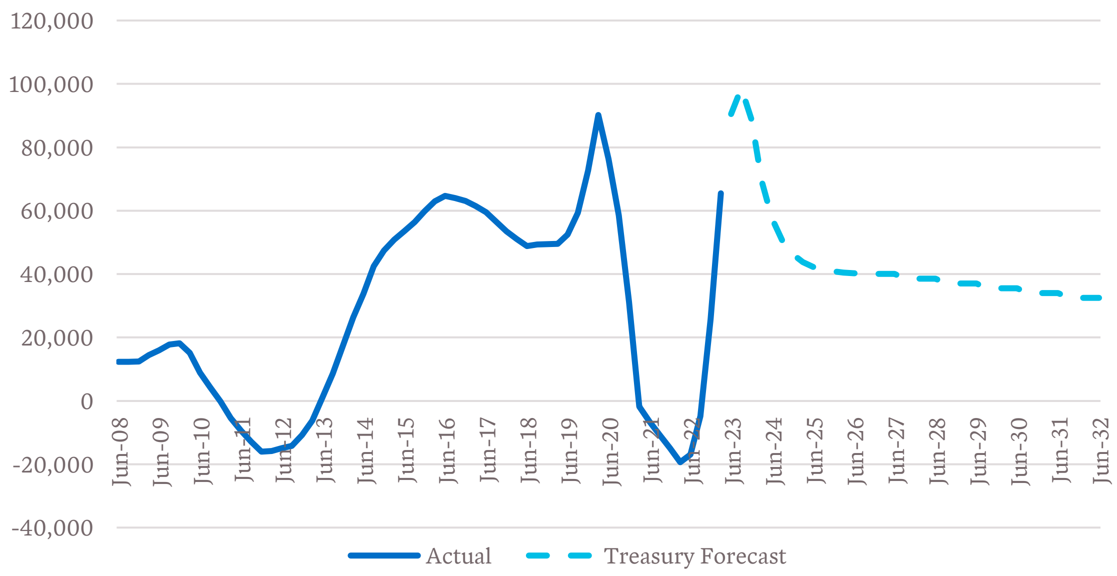
(a) The discontinuity between the previous projection and actual is due to revisions in recent periods in the underlying data.

6.1.5 Net migration

Net migration assumptions are required to calculate how many people need to be added to the population in each future quarter in respect of net migration. More detail on our approach to net migration is included in Section 8.3.2.

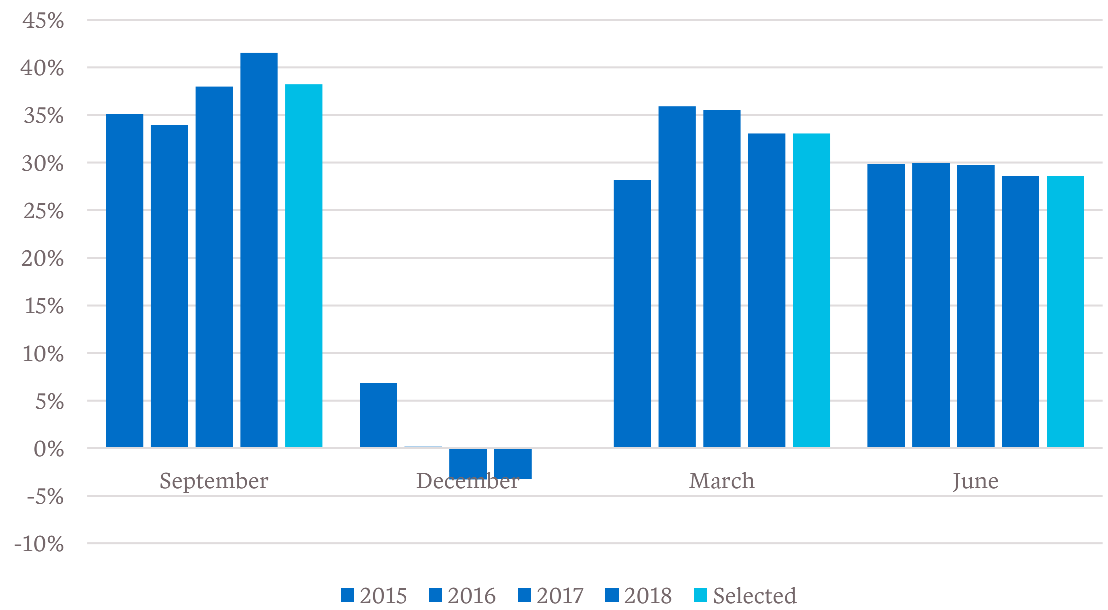
We have used Treasury forecasts in determining our assumptions for net migration numbers. They expect net migration numbers to peak around 99,000 in September 2023 reducing to 40,000 by June 2027 as post-Covid border movements stabilise, shown in Figure 6.10.

Figure 6.10 – Annualised net migration



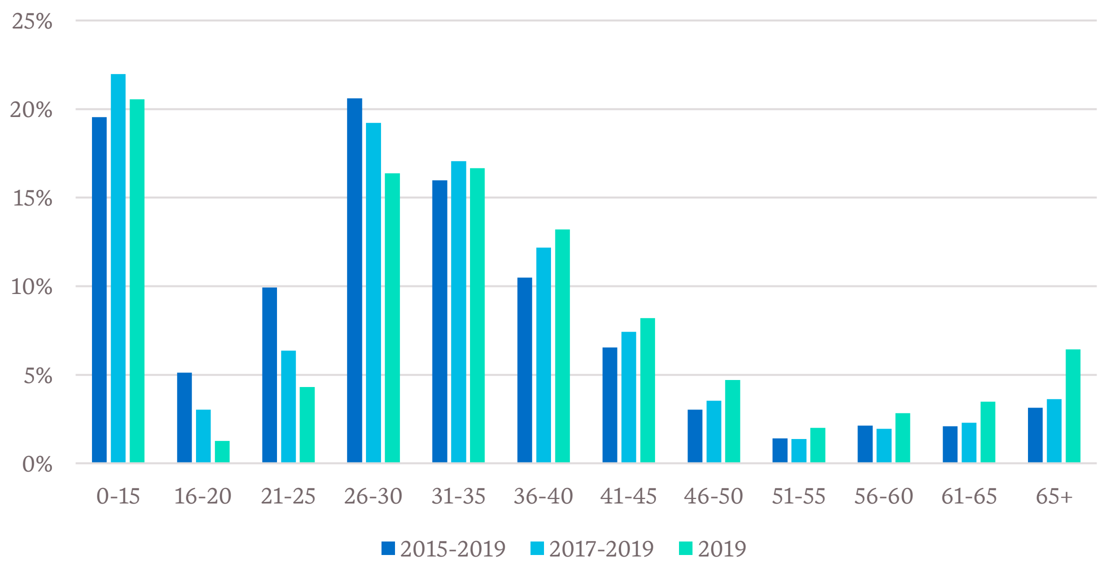
In order to project migration in our model we have converted these rolling annual averages into quarterly amounts using seasonality factors. These factors have been derived inside the IDI using the average of the movements into/out of our population (excluding births and deaths) in the four years to December 2018. We have used these four years as this is the period over which net migration has been positive, consistent with Treasury forecasts for the future. This period was also chosen so to not reflect any short-term impacts due to COVID-19. A comparison of the assumptions with the prior experience is shown in Figure 6.11.

Figure 6.11 – Proportion of net migration by quarter



Assumptions have been made by age group, with these age groups being prior to 16th birthday, 65 or older and 5-year age bands in between. These groups should be wide enough so that any volatility is smoothed out while being small enough to produce sensible populations overall. The age mixes over the past 1, 3 and 5 years of migration are shown in Figure 6.12.

Figure 6.12 Proportion of net migration by age band



We have used the average over the last five years to September 2019 for our age distribution assumptions.

Net migration volumes also vary by ethnicity and region. However, we have not incorporated this into our assumptions because it would result in very small buckets (i.e. by quarter, by age, by ethnicity, by region) and would greatly add to the computational load of the projections. Differences by ethnicity and region are instead handled by sampling these characteristics from representative populations of migrants.

The assumptions for proportion of net migration by age band do not incorporate experience beyond September 2019 due to the uncertainty in net migration statistics introduced by COVID in 2020.

6.2 General assumptions

6.2.1 Approach for setting assumptions

We use our transition and payment models to understand how emerging experience differs from what was forecast. We conduct analysis, including splitting out the impact of cyclical changes, analysis of known changes such as policy and operational changes, and consultations with MSD to give further insight into the nature of these changes. This informs a judgement about the extent to which emerging experience is likely to continue.

6.2.2 Policy reform

It is particularly hard to set assumptions relating to government service use in the presence of reforms, especially where reforms are likely to have behavioural effects that impact the way people interact with government services. Where this is the case, there is more uncertainty in estimated outcomes and a greater need to understand the sensitivity of assumptions relating to behaviours that may be impacted by reform.

Our models deliberately take a 'status quo' approach to the projection. Thus, we have not allowed for any planned future policy changes that could affect people's future outcomes and associated cash flows. It is highly unlikely that there will be no significant policy changes over the duration of the projection (50-100 years). By their nature, future reforms would be expected to affect people's outcomes with respect to government service use.

We regard this 'status quo' approach as an important feature of the projection. Setting a baseline allows us to measure the effect of future policy and operational changes as they emerge.

6.2.3 Refitting of models

The approach to refitting models was the same as that used for the 2022 model. This meant not all models were refit for 2023. Instead, as an initial step we created documents of actual versus expected charts for all models comparing the experience up to March 2023 with projections from the previous models. The previous models were based on data up to June 2022, March 2021, December 2019 or September 2018 depending on when they were last refit.

Each of these documents was then analysed to determine if that model needed refitting. Reasons for refitting included models where:

- Experience had diverged since they were last refit, and we expected this experience to be permanent
- Experience had diverged, likely due to COVID, but had not yet returned to pre-COVID levels
- Changes to the underlying data meant the existing model was a poor fit, e.g. changes to the source data for corrections variables.

6.2.4 Quarterly format

Our approach for assigning benefit and public housing states is discussed in Section 5.6.4 and Section 5.6.8.

The quarterly definition of state for both welfare and housing tends to give more stability to beneficiary numbers over time, which is useful for long-term projection.

6.2.5 Unemployment rate as a proxy for the economy

There are inherent limitations to the use of the unemployment rate as a proxy for the economy's impact on the benefit system. For example, the impact of the global financial crisis in 2007 and 2008 was greater than our modelled sensitivity would predict; many other related economic events were occurring simultaneously that compounded the impact on the benefit system. Other important economic variables include participation rates, underemployment rates, short- and long-term interest rates, credit growth, consumer spending, and business investment.

In particular, we note that the employment rate would also be a credible indicator of the labour market. It is highly correlated with the unemployment rate and can be more predictive in some cases where the unemployment rate is low. However, we have previously found that it is less sensitive in downturns compared to the unemployment rate, which may underestimate the expected influx of new clients.

We have previously examined the possibility of extending the modelling of economic variables to include other drivers, including the employment rate. However, we have found that this is difficult from a theoretical perspective (which indicators to include?) as well as a practical perspective (how to allocate signal between multiple correlated indicators?). For this reason, as per previous reports, we have chosen the unemployment rate as a strong single indicator.

6.2.6 Adjustments for the effects of COVID-19

Our approach to allowing for COVID was retained from the 2021 and 2022 models. Models were refit on data up to 31 March 2023. In many cases, this was enough time to see COVID effects appear and diminish. COVID indicator terms were added to the models to allow us to capture and run down these observed effects over time. The process was as follows:

1. For each model, use judgement and actual vs expected charts to determine if deviation of actual experience from the previous model's projections was related to COVID.
2. If the behaviour was determined to be COVID related and:

    a. Had not returned to pre-COVID levels by March 2023: refit the model using COVID indicator terms. The COVID indicator was set to 1 for the relevant period between 1 January 2020 and 31 March 2023 and interacted with other terms in the model to capture the effect. In some cases, it was interacted with particular subpopulations within a model, e.g. for entry to welfare, people working in particular industries were more impacted than those working in others.

b. Had returned to pre-COVID levels by March 2023: refit the model using time based indicator terms to back out any COVID related effects. This meant these effects would not be carried forward in the projection.

3. In the projection, we ran down the COVID indicator terms for each model from 1 to 0 over time based on how each effect was expected to diminish. The run-off was based on discussions with MSD and Treasury's unemployment forecasts.
4. Some further adjustments were required to transition rate models. This occurred in cases where the recovery from the economic effects of COVID had a significant effect on transition rates. In some cases, this recovery period had not had time to be reflected in the period the models were fit on, such as transitions out of JS-HCD which were affected by the resumption of the medical recertification process in January 2022. Additionally, the short-term unemployment forecasts from PREFU23 project higher unemployment rates in the near term. The unemployment rate is forecast to peak in 2025 before returning to the long-term average. High forecast unemployment rates result in many people being projected to move onto benefit in the next few years. Without adjustment to their transition rates these people would be projected to remain on benefit for longer periods due to their benefit history. Therefore, adjustments were made to prevent large accumulations of beneficiaries in the long term. This included adjustments to:

a. the transition rate off JS-WR and JS-HCD, which had to be increased after the unemployment peak to allow for a faster run-off than would occur with a stable unemployment rate. Stronger adjustments were applied to those with short benefit histories as these people are more responsive to economic changes and therefore more likely to exit benefit.

b. Also, while the economy is forecast to improve after 2024 and people begin to exit JS-WR and JS-HCD a higher proportion of these people will exit the system entirely rather than moving onto other benefits. Therefore, the exit rates from benefit were increased to manage the flow of people out of the system.

c. Many people leaving main benefits, but not leaving the system entirely, may receive supplementary only benefits (at least temporarily). The entry rates to supplementary benefits were increased temporarily to accommodate this phenomenon before returning to pre COVID-19 levels.

d. In our transition models, people who have some benefit history are much more likely to receive a benefit in the future. The higher number of people receiving a benefit over the short-term results in more people with benefit history. This in turn results in an increase in the number of people entering benefit over the long run. Benefit entry rates were adjusted to ensure that in aggregate the long-term entry rate is similar to historical experience prior to COVID-19.

e. During the higher forecast unemployment period, more people are likely to remain on benefits. Therefore, the transition rates from SPS and JS-HCD had to be reduced to model this effect. The transition rates from JWR did not require any adjustment during this period as the impact of the high unemployment rate was already fully captured by the model.

f. Also, more people are likely to require assistance with their housing costs during the higher forecast unemployment period. Therefore, transitions to receiving accomodation supplements were increased until the unemployment rate returned to the long term average.

g. Benefit Forecasts prepared by MSD's Forecasting and Costing team were used to help guide the level of adjustments.

6.2.7 Treatment of partners

Some benefits depend on relationship status and there are cases where both partners are supported by benefit. In theory, it would be possible to value couples as a unit as their future lifetime benefit payments are likely to be dependent. However, in the projection we have treated all clients individually, so that a primary client and their partner have separate future duration and payment estimates.

One practical implication for this approach is that much of MSD's reporting of welfare is based around counting couples as single units. Thus, there will be some differences in attempting to reconcile numbers in this report to other published numbers.

6.3 Sensitivity analysis

Note: The sensitivity analysis was last completed in 2019 and the following figures presented in this report are based on 2019 SOM results. They are included in this report to give a broad indication of the level of sensitivity in SOM results to changes in assumptions. It is reasonable to assume that the level of sensitivity in the 2023 SOM results is similar.

The model attempts to estimate a range of outcomes for people over a long-time horizon. Doing so involves making many assumptions and predictions about the future, most of which will turn out to be incorrect in hindsight; it is impossible to know exactly how the economy, inflation and behaviours will evolve. We have attempted to choose assumptions so that the resulting projection is a central estimate. Loosely speaking, we believe that the total projected amounts of different outcomes and associated cashflows are just as likely to be too high as too low.

We attempt to understand, convey, and to the extent possible, quantify this uncertainty in several ways. First, we discuss how sensitive the projection is to various model assumptions regarding key drivers. Sensitivity analysis clarifies the relationship between key drivers and the results; for example, if the unemployment rate were to track higher than the forecast level. Benefit dynamics are particularly sensitive to labour market conditions, so we also consider alternative economic scenarios to help understand the role of labour market uncertainty. Housing dynamics are particularly sensitive to rental market rates, so we consider variations on how these will change in the future.

6.3.1 Model sensitivities

6.3.1.1 Sensitivity to inflation and investment return assumptions

Some assumptions in the model are explicit, and the degree to which the adopted assumption has an impact on the results can be measured by sensitivity tests. Such assumptions include inflation and discount rates as well as transition probability assumptions.

Assumed inflation rates affect a range of cashflows in the model, most notably benefit payments. Investment returns affect how much interest is earned on a notional sum set aside today. A higher rate of return means that less money needs to be set aside today, lowering future cost in today's terms. The inflation and investment return assumptions in the model are explicit. Both these rates are set according to NZ Treasury accounting assumptions (see section 6.1). Both these assumptions can change significantly from year to year, and so form part of the annual change in estimates of cashflows.

Table 6.1 shows that a 1% increase in CPI inflation would increase the 2019 estimate of future lifetime benefit payments to by 10.6% and increase the estimate of future lifetime housing payments by 9.2%. The % increase for housing payments is lesser than for benefit payments because a higher CPI inflation rate increases future benefit payments for tenants receiving a benefit, thus reducing their future IRRS payments.

A 1% increase in the discount rate would decrease the 2019 estimate of future lifetime benefit payments by 14.5% and decrease the estimate of future lifetime housing payments by 19.5%. The timing of future payments affects the degree of impact on the estimates of future lifetime payments (or more accurately termed the present value of future lifetime payments). The further a payment is in the future, the greater the impact of a uniform change to discount rates. Housing payments have a higher weighted average duration than benefit payments, so the effect of a 1% increase in discount rates is higher in percentage terms. Often, but not always, inflation and discount rates change in the same direction. In which case their effect on the estimates of future lifetimes payments will offset each other to some degree.

Table 6.1 – 2019 Total future lifetime payments: sensitivity to changes in assumed inflation and discount rates

Scenario	Future lifetime benefit payments ($bn)	Change ($bn)	Change (%)	Future lifetime housing payments ($bn)	Change ($bn)	Change (%)
Base	142.6			115.4		
CPI Inflation +1%	157.7	15.1	**10.6%**	126.1	10.7	**9.2%**
AWE Inflation +1%	169.6	27.0	**19.0%**	137.4	22.0	**19.0%**
Discount rate +1%	122.0	-20.6	**-14.5%**	92.9	-22.5	**-19.5%**
Market rent growth +1%	142.5	-0.1	**-0.0%**	137.8	22.4	**19.4%**

Note: numbers may not sum exactly due to rounding used for displaying results

6.3.1.2 Sensitivity to transition dynamics

The transition model assumptions affect how people are forecast to move through the benefit and public housing systems each quarter. These have a significant impact on the projected incidence of benefit receipt and public housing use.

Table 6.2 shows the sensitivities of the 2019 results to changes in the probability of three key transition dynamics. Using the first case as an example, a 5% increase in a transition rate means that a client with a 20% probability of leaving JS-WR in a quarter is changed to 21% (=20% x 1.05).

We see that, of the transitions listed, the General Population entry rate causes the largest impact on the future lifetime benefit and housing payments.

Table 6.2 – 2019 Total future lifetime payments: sensitivity to changes in key transition dynamics

Scenario	Future lifetime benefit payments ($bn)	Change ($bn)	Change (%)	Future lifetime housing payments ($bn)	Change ($bn)	Change (%)
Base	142.6			115.4		
JS-WR leave rate +5%	141.9	-0.7	**-0.5%**	115.1	-0.3	**-0.3%**
PH leave rate +5%	142.8	0.2	**0.2%**	115.2	-0.3	**-0.2%**
Entry rate for first time on ben +5%	143.7	1.1	**0.8%**	115.8	0.4	**0.3%**

Note: numbers may not sum exactly due to rounding used for displaying results

6.3.1.3 Sensitivity to the unemployment rate assumption

Table 6.3 shows the sensitivities of the 2019 results to increasing the unemployment rate assumption by 1%.

Table 6.3 – Total future lifetime payments: sensitivity to changes in the assumed unemployment rates

Scenario	Future lifetime benefit payments ($bn)	Change ($bn)	Change (%)	Future lifetime housing payments ($bn)	Change ($bn)	Change (%)
Base	142.6			115.4		
UE rate + 1%	157.6	15.1	**10.6%**	120.3	4.9	**4.2%**

Note: numbers may not sum exactly due to rounding used for displaying results

6.3.1.4 Simulation variability

Our projection models are simulation based, in that we use the models to simulate a person's future outcomes multiple times and average the results.

These pathways carry a large amount of variability. Given a group of individuals with similar characteristics, we would expect them to experience different outcomes in the future. The simulation process adds variability to our results. We reduce this simulation variability in our estimates by:

- Running five independent simulations for each individual and taking the average.
- Aggregating many individual results. In any single simulation some individuals will be higher than their true average, some lower, but these will tend to balance across the projection.

In theory, it would be possible to generate a series of 'unlucky' simulations, which biased the estimate too high or low. However, simulation error is one of the smallest sources of uncertainty in the projection.

6.4 Data assumptions

6.4.1 Ethnicity

Our approach for assigning ethnicity is discussed in Section 5.6.2.1.

6.4.2 Location

MSD has 11 regions that it uses to manage its services. These are shown in Figure 6.13.

However, for housing purposes these regions are not sufficiently granular to capture differences in housing-related costs, demand and supply. Therefore, the model projects at a Territorial Local Authority (TLA) level (65 of them, excluding Auckland). Auckland is a single TLA, so we further split this into the 20 local boards. The mapping of TLAs and Auckland boards to region is shown in Table 6.4. Note that these groupings are not entirely exact; some TLAs straddle more than one Work and Income region. In this case, the TLA is placed in the Work and Income region that the majority of it falls into.

Figure 6.13 – Work and income regions

Table 6.4 – List of TLAs and Boards plus associated Work & Income region

Region	TLA/Board	Region	TLA/Board	Region	TLA/Board
Northland	Far North District	Central	Horowhenua District	Southern	Invercargill City
Northland	Kaipara District	Central	Kapiti Coast District	Southern	Mackenzie District
Northland	Whangarei District	Central	Manawatu District	Southern	Queenstown-Lakes District
Waikato	Hamilton City	Central	Masterton District	Southern	Southland District
Waikato	Hauraki District	Central	Palmerston North City	Southern	Timaru District
Waikato	Matamata-Piako District	Central	Rangitikei District	Southern	Waimate District
Waikato	Thames-Coromandel District	Central	Carterton District	Southern	Waitaki District
Waikato	Waikato District	Central	South Wairarapa District	Auckland	Albert-Eden Local Board Area
Waikato	Waipa District	Central	Tararua District	Auckland	Devonport-Takapuna Local Board Area
Bay of Plenty	Kawerau District	Wellington	Lower Hutt City	Auckland	Franklin Local Board Area
Bay of Plenty	Opotiki District	Wellington	Porirua City	Auckland	Henderson-Massey Local Board Area
Bay of Plenty	Rotorua District	Wellington	Upper Hutt City	Auckland	Hibiscus and Bays Local Board Area
Bay of Plenty	South Waikato District	Wellington	Wellington City	Auckland	Howick Local Board Area
Bay of Plenty	Taupo District	Nelson	Buller District	Auckland	Kaipātiki Local Board Area
Bay of Plenty	Tauranga City	Nelson	Grey District	Auckland	Māngere-Ōtāhuhu Local Board Area
Bay of Plenty	Western Bay of Plenty District	Nelson	Kaikoura District	Auckland	Manurewa Local Board Area
Bay of Plenty	Whakatane District	Nelson	Marlborough District	Auckland	Maungakiekie-Tāmaki Local Board Area
East Coast	Central Hawke's Bay District	Nelson	Nelson City	Auckland	Ōrākei Local Board Area
East Coast	Gisborne District	Nelson	Tasman District	Auckland	Ōtara-Papatoetoe Local Board Area
East Coast	Hastings District	Nelson	Westland District	Auckland	Papakura Local Board Area
East Coast	Napier City	Canterbury	Ashburton District	Auckland	Puketāpapa Local Board Area
East Coast	Wairoa District	Canterbury	Christchurch City	Auckland	Rodney Local Board Area
Taranaki	New Plymouth District	Canterbury	Hurunui District	Auckland	Upper Harbour Local Board Area
Taranaki	Otorohanga District	Canterbury	Selwyn District	Auckland	Waiheke Local Board Area
Taranaki	Ruapehu District	Canterbury	Waimakariri District	Auckland	Waitākere Ranges Local Board Area
Taranaki	South Taranaki District	Southern	Central Otago District	Auckland	Waitematā Local Board Area
Taranaki	Stratford District	Southern	Clutha District	Auckland	Whau Local Board Area
Taranaki	Waitomo District	Southern	Dunedin City		
Taranaki	Whanganui District	Southern	Gore District		

The more granular TLA/Auckland local board is an important predictor for housing supply and demand as well as market rental levels and thus IRRS and AS levels.

6.5 Changes since previous report

None, beyond updating assumptions as described.

C

Projection methodology

7 Projection

In this section we:

- Set out the components of the projection framework
- Describe the quarterly update of the projection
- Discuss the type of results available.

7.1 Introduction

The 2023 social modelling results are drawn from a projection of a large number of components:

- The evolution of the current and future resident adult populations in New Zealand over a ten-year period
- Entry into the future resident population both from children ageing in and net migration
- Individual characteristics that are predictive of welfare and housing utilisation
- Individual pathways through the welfare system
- Individual touchpoints on public housing and accommodation support
- Individual income amounts
- Other social sector individual characteristics.

The resulting meta-projection provides insights on social outcomes among different cohorts. Fiscal-related information such as expected future cost is available, together with more socially orientated information such as time supported by benefits and cohorts with high likelihood of certain outcomes (and thus candidates for intervention policies).

Given the sheer number of components, the projection is necessarily a complex application. This section provides a high-level overview of the steps in the projection. It also examines the technical considerations in model development for the various components of the projection.

In this section we provide details on the following:

- Projection components
- An overview of the steps in the projection for each quarter
- Output data.

Following sections drill down into the specific details underlying each of the main components:

- Economic assumptions (Section 6)
- Population model (Section 8)
- Benefit system models (Section 9)
- Public housing models (Section 10)
- Income models (Section 11)
- Other models (Section 12).

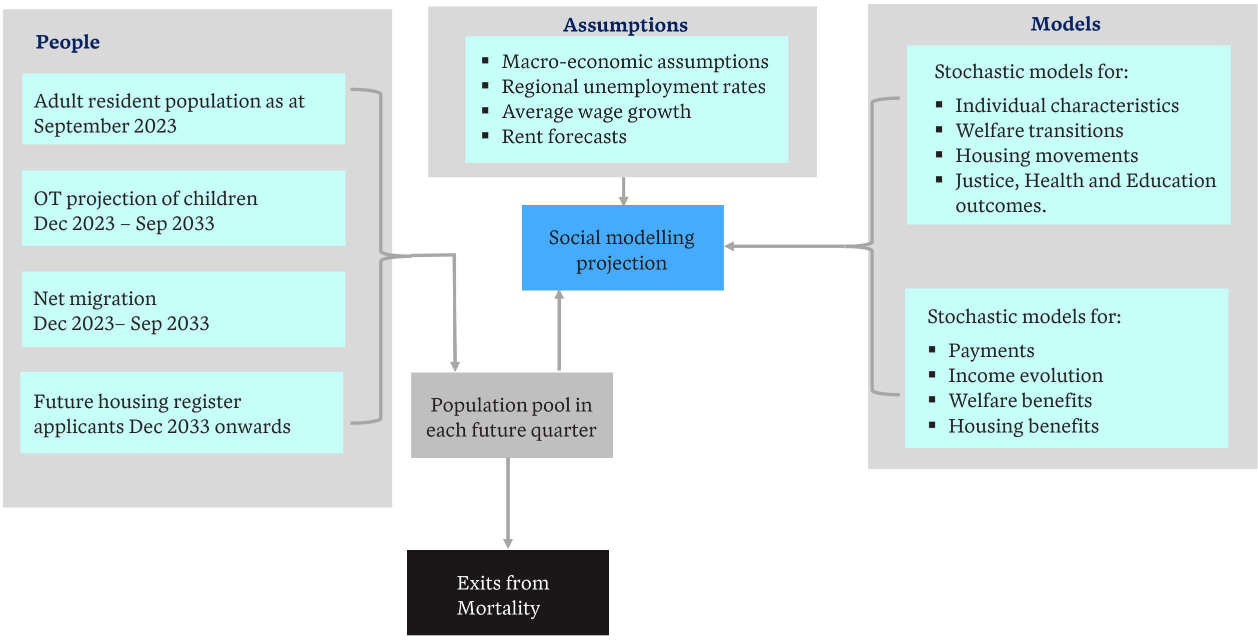
7.2 Projection framework

An overview of the components is provided in Figure 7.1. There are three main components:

- People

- Models
- Assumptions

Figure 7.1 – Projection components



7.2.1 People

The population in each future quarter is a critical part of the projection. The starting point for this is the current adult resident population as at September 2023. The model is required to cover the population in the first ten years; thus, as well as evolution of the September 2023 population, the projection framework needs to capture new entrants into the adult resident population. New entrants come from two sources:

- Children aged under sixteen as at September 2023 ageing into the adult population (drawn from the children's model projections) for the first ten years
- Net migration of adults and children to New Zealand up to September 2033.

Explicit population exits arise from mortality only since the projection considers net migration rather than separate immigration and emigration.

After September 2033, the projection no longer captures the full resident adult population. However, for housing supply and demand dynamics, it is necessary to allow for future housing register applicants from those not captured in the estimated adult resident population as at September 2033.

All individual characteristics must attach to each person, both in the starting cohort and those who enter in the future.

Finally, mortality assumptions are required to ensure the correct ageing of the population cohorts.

Further details on the population related component are in Section 8.

7.2.2 Models

A large number of models under-pin the welfare and housing modelling and provide a mechanism for:

- Forecasting transitions through welfare states
- Forecasting public housing usage
- Estimating future income levels

- Estimating future welfare and public-housing related costs
- Updating individual characteristics and other service usage.

Projection modifications are sometimes made to specific models when validating projection results. This is done:

- To address any areas where the results are not quite as we expect
- On direction from the Ministry, to broadly align estimates of future numbers of people on main benefits with the Ministry's internal forecasts.

7.2.3 Assumptions

Finally, assumptions around the macro-economic environment are a critical part of the projections. Economic data used within the projections includes:

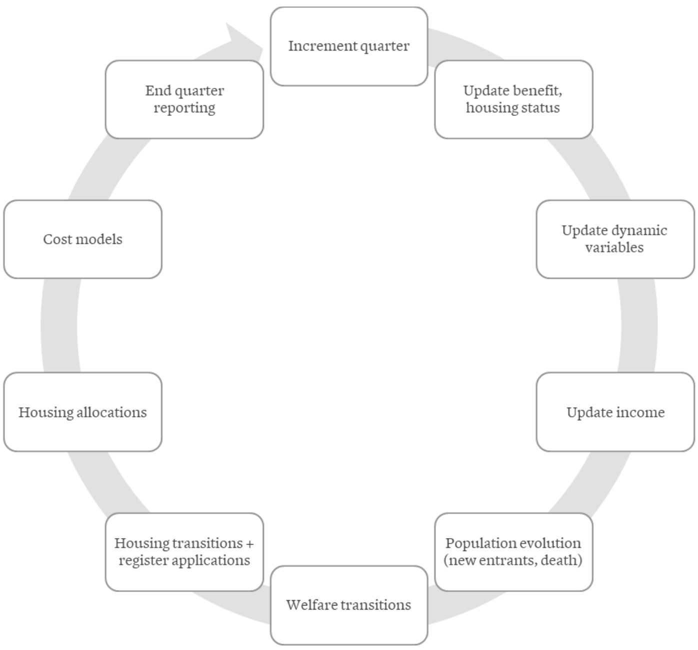
- Regional unemployment rates
- Average wage growth
- Future rental costs
- Net migration.

Macroeconomic assumptions have been discussed in Section 6.

7.3 The quarterly projection process

The social outcomes model aims to provide a snapshot of the adult resident population each quarter. Thus, the projection runs quarter by quarter and consists of the steps set out in Figure 7.2.

Figure 7.2 – Quarterly projection steps



One key thing to note about the projection process is that we forecast the welfare and housing states in the following quarter and not those states in the current quarter. This means that we use individual characteristics as in the current quarter, as well as the current housing and welfare states to determine the expected states in the next quarter.

In more detail the steps are:

1. Start with the adult resident population as at the end of the previous quarter, increment the quarter counter and update their benefit and housing status to be the forecasts of these states made in the previous quarter.
2. Update other characteristics of this population (e.g. age, incapacity, children etc.). Some quantities update in a known fashion (e.g. age, benefit history which is deterministic given the current benefit state) while others (birth of a new child, change of district etc.) must be modelled.
3. Forecast income levels and industry based on characteristics and previous income levels.
4. Allow for register movements – including new and transfer applications.
5. Allow for population new entrants – young people ageing in, positive net migration.
6. Allow for deaths through mortality transitions.
7. Project the welfare state in the following quarter for everyone under the retirement age.
8. Allow for housing movements – exits from housing, register applications, Accommodation Supplement receipt.
9. Allocate public houses to those on the housing register.
10. Apply the cost models to produce estimates of welfare, superannuation and housing costs.
11. Output the end of quarter results, which may include using some overlay models to forecast end quarter benefit and housing state.
12. Increment the quarter and forecast the next quarter, by starting again at step 1.

7.4 Results

The final stage in each quarter of the projection is to output the results. At their most granular level, results are available for each person and for each quarter. Typically, the information recorded includes:

- Personal identifiers
- Region
- Benefit type
- Housing status and related housing information
- Welfare and housing payments
- Income
- Industry
- Health
- Offences leading to Police proceedings and corrections.

However, results are usually reported at a summarised level since individual results reflect average pathways rather than a detailed specific forecast for an individual. Thus, results will generally be presented for subgroups of interest.

7.5 Changes since previous report

None.

8 Population model

In this section we describe how the population being modelled is defined and how this is derived from the available data. In particular, we describe:

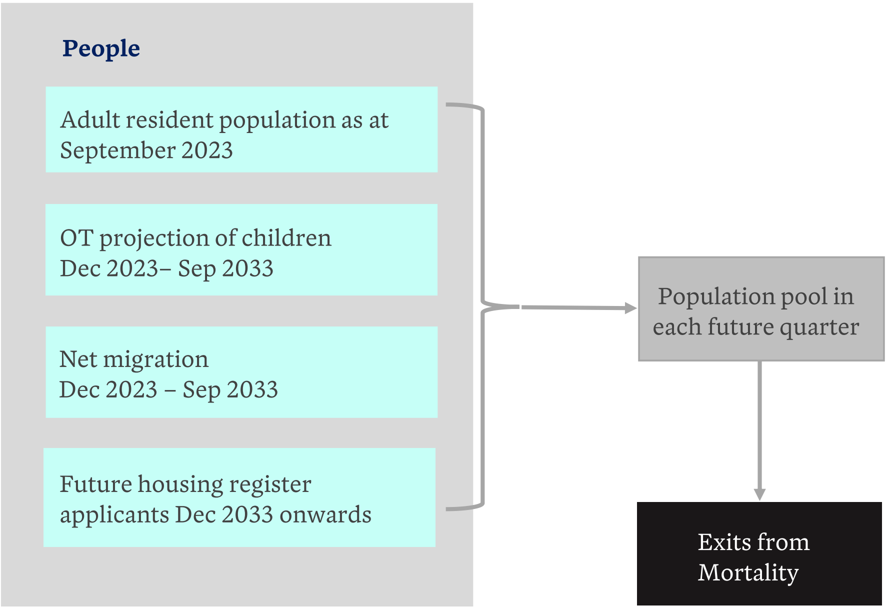
- The starting population as at 30 September 2023
- Future entrants to the starting population
- Mortality
- How we have reduced computational load by grouping the general population into 'buckets' and how we have determined these bucket sizes.

8.1 Introduction

The 2023 model takes a whole of population approach for the first ten years. The starting projection cohort contains all resident adults in New Zealand as at September 2023. Thereafter, the projection allows for the evolution of the population through new entrants (children ageing in, immigration) and exits (mortality, emigration).

Figure 8.1 summarises the various components in the evolution of the population over the first 10 years, and thereafter.

Figure 8.1 – Population evolution



This section provides further technical details on the data, models and processes underpinning population evolution.

8.2 Adult resident population as at 30 September 2023

Details on the derivation of the adult resident population is included in Section 5.6.1.

8.3 Future entrants to the projection

The model projects the full resident population for the first 10 years. Thus, new entrants to the population over that time period are required inputs into the model. There are two sources for these new entrants:

children ageing into the resident adult population, i.e. turning 16 years old in the projection quarter and net migration.

After the tenth year, the projection continues quarter by quarter to estimate the future lifetime pathways and costs of all individuals in the projection at that point. Results for individuals who enter the population after ten years are out of scope. However, to ensure the proper operation of the housing supply and demand model, it is necessary to continue adding future register applications from new entrants to the population.

Each of these components is described below.

8.3.1 Child entrants from the children's model

Child entrants are the main source of future entrants to the population and are sourced from Oranga Tamariki's children's model. The children's model itself is a micro-simulation of the child population in New Zealand, so is available at the granular level of one row per child turning 16 in each of the forty quarters from December 2023 to September 2033.

Further details on the output of the children's model and how this is linked into our projection model are given in Section 5.5.

8.3.2 Future entrants from migration

The model considers net migration rather than separate assumptions for emigration and immigration. For positive forecast net migration, the effect of this is to add new entrants for different population cohorts. We do not allow for negative net migration.

The benefits of using a net migration approach rather than separate assumptions include:

- The approach is significantly simpler computationally. Having separate immigration and emigration assumptions means the model would need to keep track of those who have emigrated in case they are to later return to New Zealand. This would expand the population significantly.
- Separate assumptions would also require a model to be built to determine who is most likely to emigrate rather than using sampling as we have. Getting the wrong characteristics of those who are likely to emigrate could have a big impact on the model.
- Treasury forecasts migration only on a net basis.
- Net migration is forecast to be positive in future years from Treasury and Statistics NZ. This means we do not have to remove individuals from the population. The additional migrants can be given consistent characteristics to those who have migrated in the past.
- The children's model does not allow for migration in any form. Applying separate immigration/emigration assumptions to the entrants at age 16 would hence be very difficult.

There are some disadvantages to using this approach, including:

- Having net migration assumptions means we are sampling from only new entrants and can result in the distribution of people in the projected population for some segments being different than what it would be from separate assumptions. By allowing for the age distribution, we ensure we get sensible numbers in the population by age group, which is one of the key model terms. We previously noted that we identified variability by ethnicity and region as well, but control for this by sampling these characteristics from a representative population – see Section 6.1.5.
- Immigrants may be more/less likely to interact with the welfare/housing systems than emigrants. We have looked into this aspect and found that a similar number of people are supported by a benefit within a year of immigrating to New Zealand as there are people who emigrate from New Zealand within a year of being supported by a benefit.

- This approach does not allow full analysis of migrants and their service use. We see this as a minor disadvantage of the approach. Service use of those who are in the population and have not interacted with the benefit system before is much more important.

On balance, we consider the migration approach taken will not give materially different modelling results compared to more detailed approaches that explicitly model immigration and emigration.

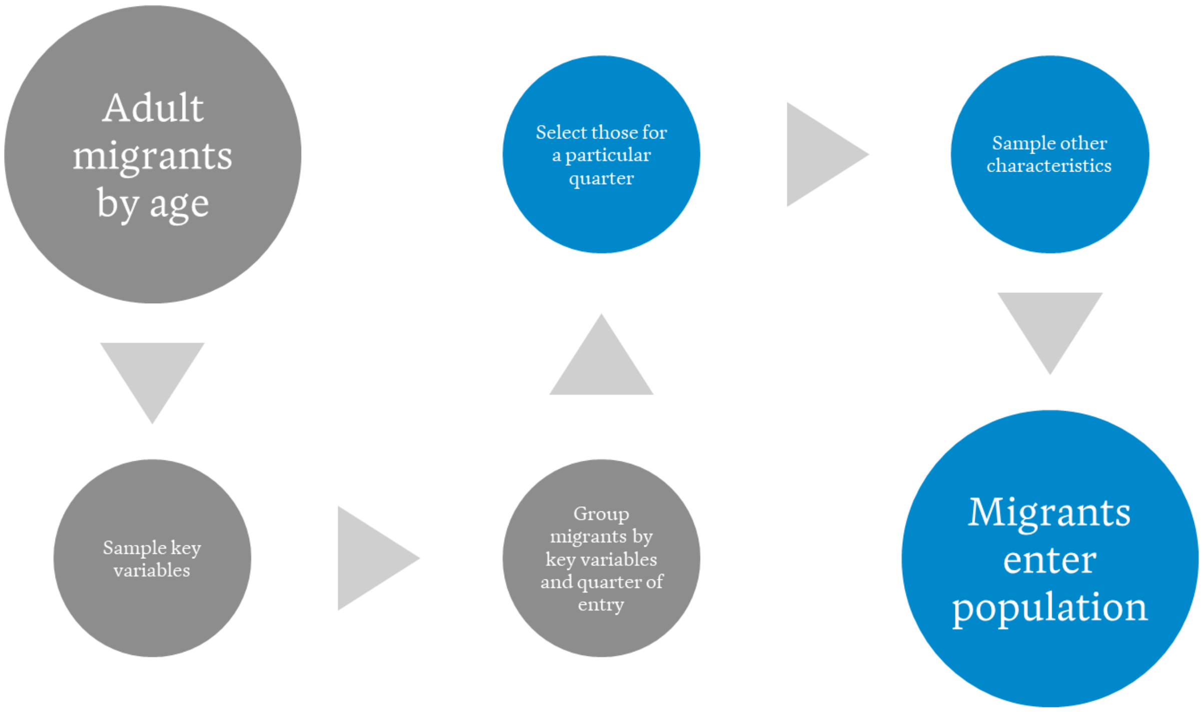
As far as the projection mechanics are concerned, people are forecast to enter the population via migration in two ways:

- Migration of adults
- Migration of children to NZ who then age into the projection when they turn 16.

8.3.2.1 Migration of adults

The process by which adult migrants are added is shown in Figure 8.2. A number of pre-processing steps take the net migration Treasury forecasts for the next ten years (where they are positive), which are available by age only, and attach key characteristics (sex, ethnicity, region, education, previous benefit or housing history) via sampling from a representative population. The resulting data set is then an input to the projection. For the first ten years in each projection migrants are added from this data set.

Figure 8.2 – Adding adult migrants



These migrants are then added to the population. As they only contain data for the key characteristics of age, sex, ethnicity, education and region, their remaining characteristics are sampled from a representative table of such characteristics, which is also stratified by age, sex, ethnicity, education and region.

8.3.2.2 Migration of children

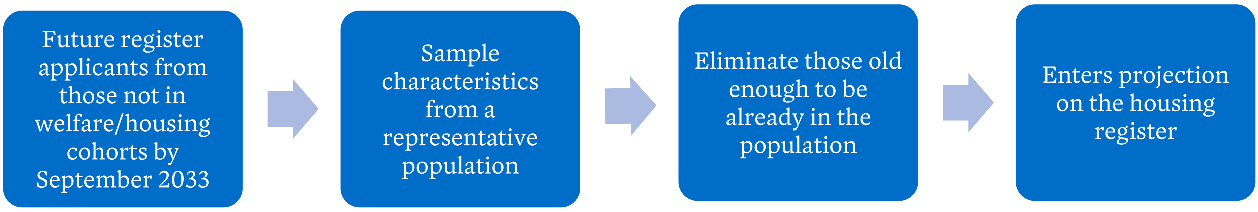
The children's model only includes those who are currently in the population with no migration allowed for. We have allowed for migration of children who subsequently age into the resident adult population by adjusting upwards the numbers forecast to enter from the children's model to account for child migrants. This adjustment is based on actual migration over the last three years.

8.3.3 Projection entrants after 10 years

The focus of the projection is on the population up to and including September 2033 and the experience of this population. Thus, future population entrants from September 2033 are not required for incorporation into results.

However, to ensure that social housing dynamics of supply and demand function as expected, it is necessary to allow for register applicants from population entrants beyond September 2033. Figure 8.3 summarises how these people are incorporated into the projection.

Figure 8.3 – Register applicants from September 2032 onwards



The following steps are used to add future register applicants:

- For each quarter, the forecast number of future applicants, not already in the housing or welfare cohorts by September 2033, is extracted from a table. This table is the same as that used in the previous report.
- Characteristics for each new entrant are sampled from a table of representative features.
- Based on these characteristics, those already in the population are eliminated. The criterion used here is based on age – only those future entrants who are too young to have been in the child cohorts from the children's model that have aged into the projection are permitted to enter the projection cohort via the housing register.

8.4 Mortality

Mortality models estimate whether someone will die in the next quarter and are applied before transition models. It is estimated via the use of GLMs that estimate the probability of death.

Population mortality rates have been sourced from Statistics NZ life tables. These are split by future quarter, age last birthday, sex and ethnicity. Stats NZ life tables have been used as an offset in the model meaning that overall, our projections will be in-line with population mortality rates. Future expected mortality improvements are implicit within the life tables as they are split by future quarter.

Models have been segmented by age, namely:

- Aged less than or equal to 65
- Aged greater than 65.

Other dynamic variables are used in the model such as benefit and housing state as well as health conditions. We have kept the number of variables used in these models to a minimum as we do not want individuals to have large fluctuations in their expected mortality rates from quarter to quarter. For the aged over 65 model, as people get older, we have phased out other variables so individual mortality rates become more closely aligned with population mortality rates.

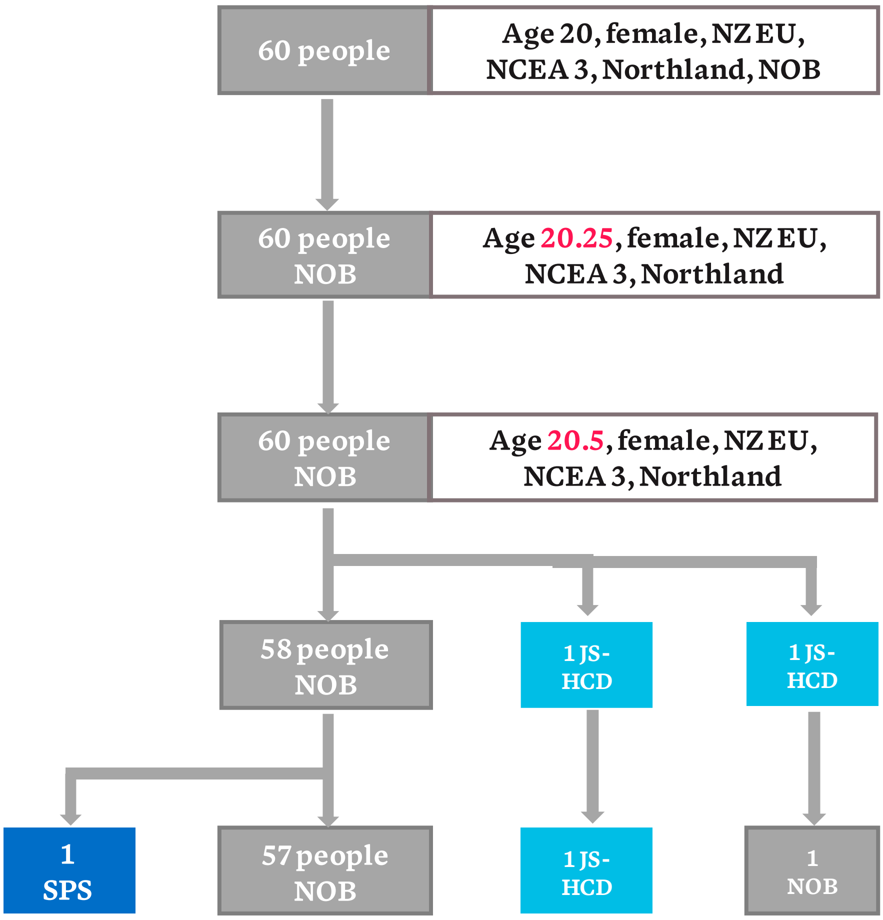
8.5 Evolution of the general population

Rather than having one row per adult in the projection, for computational reasons, we group the general population, and future entrants, into groups of a variable size (those on welfare and/or housing are identified separately as one entry per individual). We refer to these as ‘bucket’ rows. These bucket rows evolve as follows:

- Dynamic population characteristics such as health, corrections, income etc. are assumed to evolve in the same way for all members of the group.
- Housing and welfare state transitions, including mortality, operate by projecting the number of individuals from the row that will enter housing and/or welfare in the next quarter (in many cases this will be zero). Where this is non-zero, these people “break off” from the grouped entry and are allocated a new individual entry. The number of individuals in the grouped row is reduced accordingly.

Figure 8.4 gives an example of the evolution of a general population bucket row through time. For simplicity, we only show break-up of bucket rows due to welfare transitions, but the same logic applies to housing transitions (register application, receiving AS) and mortality.

Figure 8.4 – Evolution of the general population bucket rows through time



8.6 Variable bucket sizes

We have followed a rough rule of thumb that 1000 individual rows (including buckets) are required to be able to analyse results reliably. In order to be able to analyse results for general population groups of interest we have used a method in which bucket size can vary. This means that for groups of interest where the default bucket size of 60 does not produce the required 1,000 rows but there are at least 1,000 people in the starting population, we adjust the size to a level that will produce more robust results. This is particularly the case when cohorts are broken down by some of the variables that MSD are interested in such as Oranga Tamariki interaction or Corrections history.

To determine the bucket sizes, we segment the population based on 10 variables:

- Age
- Sex
- Ethnicity
- Education level (using NZQF achievement at school)
- Region
- Whether or not the person has ever received welfare or been in public housing
- Having had any mental health event in the last three years
- Having any corrections history in the last ten years
- Having any Oranga Tamariki interaction
- Having a history of intergenerational benefit receipt.

Each segment is then given a starting bucket size of 60. Various combinations of the variables are used to segment the population into chosen subgroups for which results may need to be analysed for. The combinations are a small subset of variables from each of the segmenting lists above.

Then the groups are looped through, and bucket sizes for those groups are updated in the following process:

1. Determine the number of individual rows the currently allocated bucket sizes will produce
2. If this is less than 1000, reduce the bucket size for all segments in that group to get at least 1000 rows. If there are less than 1000 people in the group and therefore getting 1000 individual rows is not possible, give all people in that group a bucket size of 1.

The groups of interest must be updated one at a time, as some of the segments belong to more than one group of interest. Changing the bucket size for all segments in one group will change the number of rows produced for another group. For example, changing the bucket size for all females with a history of mental health service use will impact the bucket sizes for all people in Auckland with a history of mental health, as the groups overlap.

Using 60 as a default size was found to be the size that provided the optimal ratio between granularity and computational time.

Overall, this variable bucket size approach results in an increase of approximately 100,000 buckets (rows) compared to the previous approach. This results in a model that takes longer to run in R and requires more memory.

8.7 Changes since previous report

None.

9 Benefit system models

This section describes the models specific to modelling benefit system-related outcomes. It covers:

- The benefit state construct
- The models that project transitions between different states
- Payment models – the modelling of the amount of benefit payment
- Welfare individual characteristic models.

A full list of the models is included.

9.1 Introduction

A key part of the modelling is the projection of individuals' pathways through the benefit system, and the associated costs. Indeed, the modelling has grown out of previous modelling (pre-2015) work on working age welfare-spend, where the population considered included only those that had received welfare support in the most recent year or were forecast to in the next five years.

With the extension of the work in 2015 to include public housing, it was necessary to extend the population to capture all those in public housing system (or had been in the most recent year) – which included some individuals who were not captured in the welfare data. Similarly, the broadening of the projection cohort to the full adult resident population in the 2018 modelling work necessarily added a large number of individuals who have never accessed the welfare and/or public housing system as well as those who had done so at least one year in the past.

Thus, there are a number of possible working-age welfare states (including not on welfare currently and never on welfare) which individuals can move between and receive payments in. These are shown in Figure 9.1. For the projection each person is assumed to be in a single benefit 'state' each quarter. In this context, 'benefit state' refers to the current main benefit received by the client, or a state of 'SUP' or 'NOB' if a client is receiving supplementary benefits only or is not supported by benefits respectively. In all there are nine possible benefit states, corresponding to seven main benefit types, supplementary-only beneficiaries and those not receiving benefits (NOB).

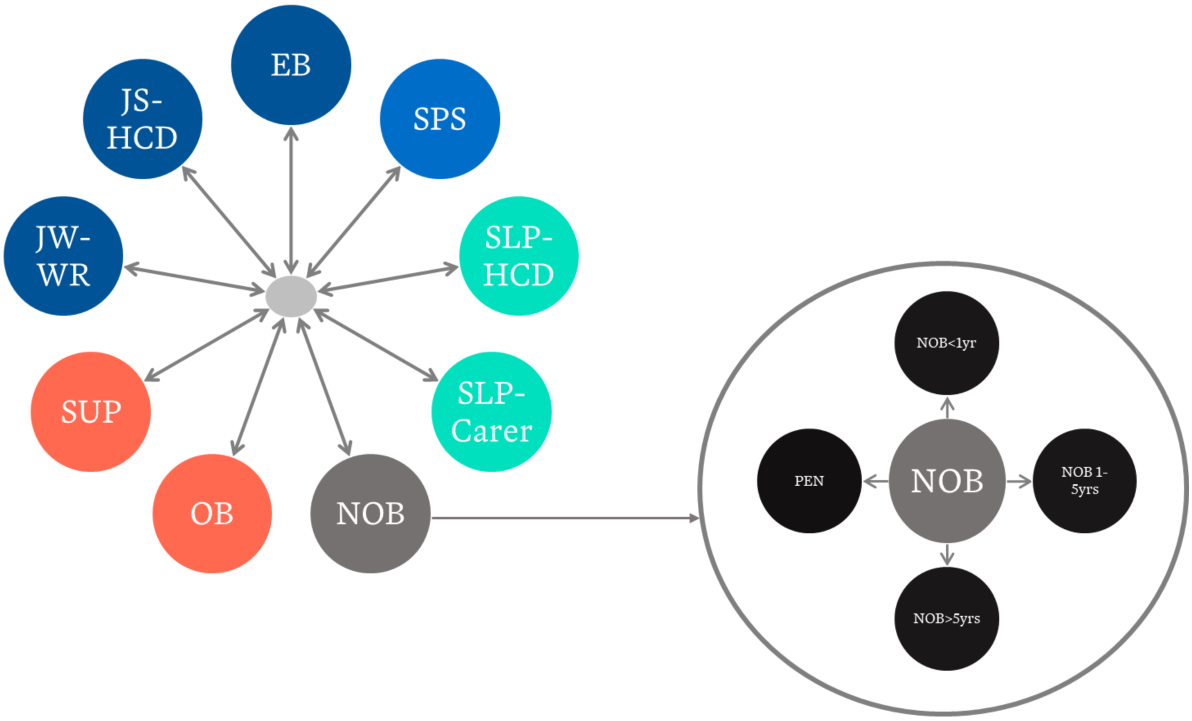
NOB consists of a number of different types of people not supported by benefits:

- Working-age people who have exited the benefit system within the last year
- Working-age people who last received benefits between one and five years ago
- Working-age people who last received a benefit more than five years ago or who never received a benefit.

Individual characteristics retained in the data permit us to identify each of these groups, which may be required for reporting and insights. They are not, however, separate welfare states in the model.

When individuals reach retirement age (currently 65 years old) they are assumed to transition into retirement which is labelled as PEN (pensioners). Once an individual is labelled as PEN, they cannot return to the welfare system (represented by the single directional arrow in Figure 9.1).

Figure 9.1 – Benefit states



There are two key types of models in the welfare portion of the projection framework:

- Transition models, which model the probability of remaining in the current state, or moving to each of the other eight states, for each quarter.
- Payment models, to calculate the average benefits received by the client given their current state.

Note that there is no model for transitions into the PEN state since individuals are forced to transition to PEN once they turn 65. Although not strictly the case in practice, we have made this assumption for simplicity.

The variables used in these models cover a range of demographics, past benefit and housing support receipt, intergenerational benefit receipt, income and interactions with other government services. The range of variables allows for detailed modelling. This is important and valuable when providing:

- Estimates of change
- Identification of protective and risk factors as well as interactions between such factors
- Analysis of subgroups.

Both these types of models are described in more detail in the remainder of this section.

9.2 Transition models

9.2.1 Overview

There are 81 (i.e. 9 x 9) different benefit state transition types. Rather than model each type of transition separately, we use a series of probability models which focus on the most probable transitions. These models are then chained together to produce the full set of transition probabilities. The most frequent benefit transitions are:

- A client remaining in their current benefit state
- A client moving from benefits to no benefits (i.e., moving into the NOB state)

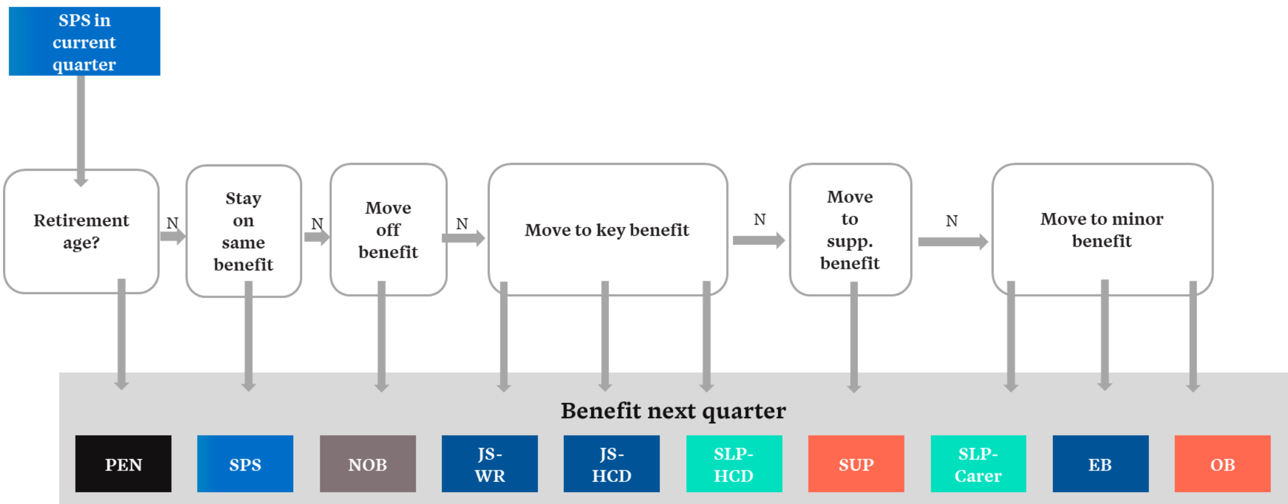
- A client moving from no benefits to benefits (moving out of the NOB state). Note that this includes returning to benefits after a period off benefits and entering benefits for the first time. The dynamics of each of these are different so are considered separately.

The process of chaining together the transition models to produce the full set of probabilities is best described using an example. Figure 9.2 shows the sequence of model applications as a flowchart to determine the benefit during the next quarter of an individual in SPS. This sequence is as follows:

- First determine if the individual has reached retirement age. If yes, then they transition to PEN.
- Next apply the transition model to determine if the individual remains in SPS. If yes, then they remain in SPS.
- If no, then apply the transition model to determine if the individual moves off benefit in the next quarter. If yes, then their benefit state becomes NOB.
- If no, then apply a multinomial model to determine their state next quarter. Possible outcomes are JS-WR, JS-HCD, SLP-HCD and Other. If one of the first three, then this becomes their benefit state in the next quarter.
- If Other, then use a transition model to determine if the individual will receive supplementary benefit types only in the following quarter. If yes, then their benefit will be SUP.
- If no, then sample from a multinomial to determine which of the remaining benefit types the individual will move to in the following quarter. This can be one of EB, SLP-Carer or OB.

Analogous sequences of chained probability models are used for the other main benefit types (JS-WR, JS-HCD and SLP-HCD).

Figure 9.2 – Transition probabilities example for recipients of SPS



The processes for transitions from NOB (both from new entrants and from returning individuals) and for transitions from the minor benefit types are similar but with some modifications. For example, the chained models for those in NOB obviously do not include a model for moving off benefit.

9.2.2 Detailed list of all models used

Table 9.1 – List of benefit transition models

Benefit state	Type	Model ID	Description
JS-WR	Logistic	jwr_tra	Probability that a client remains in JS-WR in the next quarter

Benefit state	Type	Model ID	Description
JS-WR	Logistic	jwr_nob	Probability that a client moves from JS-WR to NOB, given that they leave JS-WR
JS-WR	Multinomial	jwr_mul	Multinomial Probability of moving to JS-HCD, SLP-HCD, SPS and OTH, conditional on leaving JS-WR and not entering NOB
JS-HCD	Logistic	jhd_tra	Probability that a client remains in JS-HCD in the next quarter
JS-HCD	Logistic	jhd _nob	Probability that a client moves from JS-HCD to NOB, given that they leave JS-HCD
JS-HCD	Multinomial	jhd _mul	Multinomial Probability of moving to JS-WR, SLP-HCD, SPS and OTH, conditional on leaving JS-HCD and not entering NOB
SPS	Logistic	sps_tra	Probability that a client remains in SPS in the next quarter
SPS	Logistic	sps_nob	Probability that a client moves from SPS to NOB, given that they leave SPS
SPS	Multinomial	sps_mul	Multinomial Probability of moving to JS-WR, SLP-HCD, JS-HCD and OTH, conditional on leaving SPS and not entering NOB
SLP-HCD	Logistic	slh_tra	Probability that a client remains in SLP-HCD in the next quarter
SLP-HCD	Logistic	slh_nob	Probability that a client moves from SLP-HCD to NOB, given that they leave SLP-HCD
SLP-HCD	Multinomial	slh_mul	Multinomial Probability of moving to JS-WR, JS-HCD, SPS and OTH, conditional on leaving SLP-HCD and not entering NOB
NOB (returning)	Logistic	nob_tra	Probability that a client remains in NOB in the next quarter
NOB (returning)	Multinomial	nob_mul	Multinomial Probability of moving to JS-WR, JS-HCD, SPS, SLP-HCD and OTH, conditional on leaving NOB
Other – inwards	Logistic	oi_sup	Probability that someone entering OTH is entering SUP
Other - inwards	Multinomial	oi_mulm	Multinomial probability that someone entering OTH but not SUP enters EB, SLP-Carer or OB

Benefit state	Type	Model ID	Description
NOB (never ben)	Logistic	zzz_tra	Probability that a client remans in NOB in the next quarter, given they have never received a benefit previously.
NOB (never ben)	Multinomial	zzz_mul	Multinomial Probability of moving to JS-WR, JS-HCD, SPS, SLP-HCD and OTH, conditional on leaving NOB and never having received a benefit previously.
NOB (never ben)	Logistic	oiz_sup	Probability that someone entering OTH is entering SUP given they have never received a benefit previously.
NOB (never ben)	Multinomial	oiz_mulm	Multinomial probability that someone entering OTH but not SUP enters EB, SLP-Carer or OB given they have never received a benefit previously.
Other	Logistic	o_tra	Probability that someone in OTH remains in their current state
Other	Logistic	o_nob	Probability that someone in OTH moves to NOB, given that they leave their current state
Other	Logistic	o_key	Probability that someone in OTH moves to one of JS-WR, JS-HCD, SPS or SLP-HCD, given that they leave their current state and do not move to NOB
Other	Multinomial	o_mulk	Multinomial probability that someone in OTH moves to either of JS-WR, JS-HCD, SPS or SLP-HCD, given that they move to one of these benefits.
Other	Multinomial	o_mul2	Multinomial probability that someone in OTH moves to either of SUP, EB, SLP-Carer or OB given that they leave their current state and do not move to NOB, JS-WR, JS-HCD, SPS or SLP-HCD.

Notes:
(a) Other (OTH) in the table refers to benefits other than the main Tier 1 benefits, i.e. SUP, EB, SLP-Carer and OB

9.3 Payment models

9.3.1 Working-age benefit payments

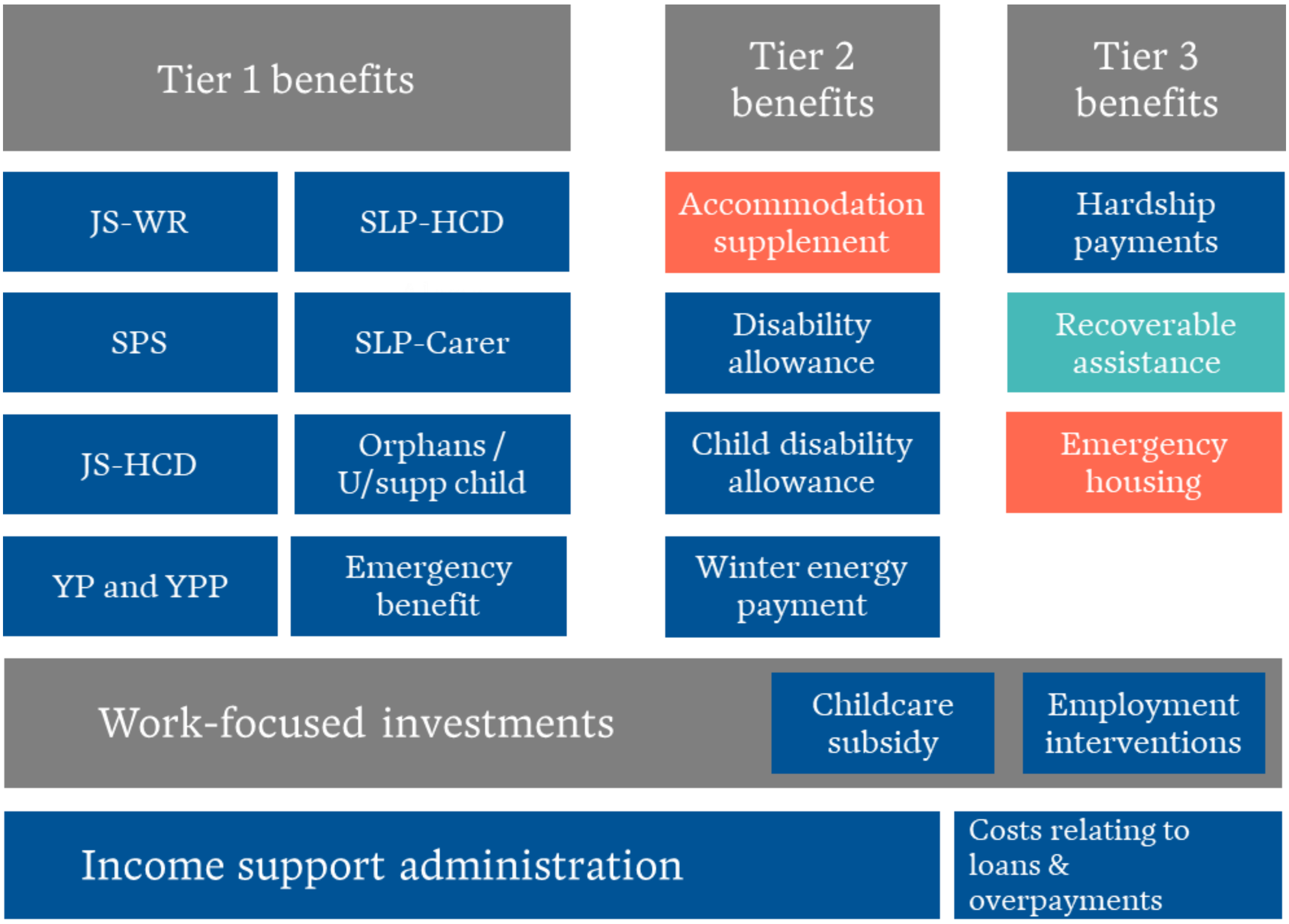
Individuals can receive several different benefit payment types simultaneously:

- A main Tier 1 payment (JS-WR, JS-HCD, EB, SPS, SLP-HCD, SLP-Carers)
- Orphans (or child living alone) Benefit (OB). Note that this may be received as a Tier 1 payment for those in OB.
- Accommodation Supplement (AS)
- Disability allowance (DA)

- Child disability allowance (CDA)
- Childcare subsidy (CCS)
- Hardship assistance (HS) and the housing-related subcomponent Temporary Additional Support (TAS)
- Employment intervention payments (EI)
- Recoverable assistance (LOA in this section)
- Emergency housing
- Winter energy payment (WEP).

Welfare payments are shown in Figure 9.3.

Figure 9.3 – Welfare payments



Notes:

(a) Accommodation supplement and emergency housing are highlighted since they are modelled explicitly as part of the housing component of the social model

(b) Recoverable assistance is partly offset by future repayments from clients

Separate payment models are required for each combination of benefit state and benefit type received while in that state. This leads to a significant number of payment models; for instance, there are eight payment models for clients in the SPS benefit state (one for the main Tier 1 benefits, OB, DA, CDA, CCS, HS, EI and Recoverable Assistance). Note we allocate all Tier 1 payments to the current benefit state. This means there is a reallocation in cases where a client receives more than one Tier 1 benefit during a quarter. The impact of this reallocation is small.

This leads to a large number of potential payments, with the relative significance of each differing greatly. Main benefits plus accommodation support make up about 90% of benefit payments to current clients, so these payment types are modelled in greater detail (AS through the housing part of the model - refer to Section 10 for full details).

9.3.2 Detailed list of all models used

Rather than including one model for every benefit type and payment combination, we have rationalised the number of models by combining payments of some of the smaller types across recipients in different benefit states. The models fitted are shown in the table below. Each of the main benefit models are fitted separately as are the larger components of Tier 2 payments (e.g. explicit modelling of DA for JS-HCD and SLP-HCD recipients and CCS for SPS recipients).

Table 9.2 – Payment models attributable to each state (excluding AS)

Payment	Type	Model ID	Description
JS-WR	Tier 1	jwr_abp	Main benefit payment for the quarter, applied to those in JS-WR state
JS-HCD	Tier 1	jhd_abp	Main benefit payment for the quarter, applied to those in JS-HCD state
SPS	Tier 1	sps_abp	Main benefit payment for the quarter, applied to those in SPS state
EB	Tier 1	emb_abp	Main benefit payment for the quarter, applied to those in EB state
SLP-HCD	Tier 1	slh_abp	Main benefit payment for the quarter, applied to those in SLP-HCD state
SLP-Carer	Tier 1	slc_abp	Main benefit payment for the quarter, applied to those in SLP-Carer state
OB/ Unsupported child	Tier 1	orp_abp	OB/Unsupported child benefit payment for the quarter, applied to those in OB state
OB/ Unsupported child	Tier 2	jwr_orp	Average OB/Unsupported child benefit payment for the quarter, applied to those in JS-WR state
OB/ Unsupported child	Tier 2	jhd_orp	Average OB/Unsupported child benefit payment for the quarter, applied to those in JS-HCD state
OB/ Unsupported child	Tier 2	sps_orp	Average OB/Unsupported child benefit payment for the quarter, applied to those in SPS state
OB/ Unsupported child	Tier 2	slh_orp	Average OB/Unsupported child benefit payment for the quarter, applied to those in SLP-HCD state
OB/ Unsupported child	Tier 2	a_orp	Average OB/Unsupported child benefit payment for the quarter, applied to those in EB or SLC-Carer states

Payment	Type	Model ID	Description
Disability Allowance	Tier 2	a_da	Average disability allowance benefit payment for the quarter paid to those in JS-WR, EB, SLP-Carer and OB
Disability Allowance	Tier 2	jhd_da	Average disability allowance benefit payment for the quarter paid to those in JS-HCD
Disability Allowance	Tier 2	sps_da	Average disability allowance benefit payment for the quarter paid to those in SPS
Disability Allowance	Tier 2	slh_da	Average disability allowance benefit payment for the quarter paid to those in SLP-HCD
Disability Allowance	Tier 2	z_da	Average disability allowance benefit payment for the quarter paid to those in SUP
Child disability allowance	Tier 2	a_cda	Average child disability allowance benefit payment for the quarter paid to those in JS-WR, JS-HCD, SLP-HCD and EB
Child disability allowance	Tier 2	sps_cda	Average child disability allowance benefit payment for the quarter paid to those in SPS
Child disability allowance	Tier 2	z_cda	Average child disability allowance benefit payment for the quarter paid to those in SLP-Carer, OB and SUP
Childcare subsidy	Tier 3	a_ccs	Average childcare subsidy benefit payment for the quarter paid to those in JS-WR, JS-HCD, SLP-HCD and EB
Childcare subsidy	Tier 3	sps_ccs	Average childcare subsidy benefit payment for the quarter paid to those in SPS
Childcare subsidy	Tier 3	z_ccs	Average childcare subsidy benefit payment for the quarter paid to those in SLP-Carer, OB and SUP
Childcare subsidy	Tier 3	nob_ccs	Average childcare subsidy benefit payment for the quarter paid to those in NOB
Hardship payment	Tier 3	jwr_hs	Average hardship benefit payment for the quarter paid to those in JS-WR
Hardship payment	Tier 3	jhd_hs	Average hardship benefit payment for the quarter paid to those in JS-HCD
Hardship payment	Tier 3	sps_hs	Average hardship benefit payment for the quarter paid to those in SPS
Hardship payment	Tier 3	slh_hs	Average hardship benefit payment for the quarter paid to those in SLP-HCD

Payment	Type	Model ID	Description
Hardship payment	Tier 3	a_hs	Average hardship benefit payment for the quarter paid to those in SLP-Carer, EB and OB
Hardship payment	Tier 3	z_hs	Average hardship benefit payment for the quarter paid to those in SUP
Hardship payment	Tier 3	nob_hs	Average hardship benefit payment for the quarter paid to those in NOB
Employment interventions	Tier 3	x_ei	Average employment interventions benefit payment for the quarter paid to those in JS-WR, SPS and EB
Employment interventions	Tier 3	a_ei	Average employment interventions benefit payment for the quarter paid to those in JS-HCD, SLP-HCD, SLP-Carer, OB and SUP
Employment interventions	Tier 3	nob_ei	Average employment interventions benefit payment for the quarter paid to those in NOB
Net Recoverable assistance	repayment	jwr_loa	Average amount of recoverable assistance received by clients in JS-WR
Net Recoverable assistance	repayment	jhd_loa	Average amount of recoverable assistance received by clients in JS-HCD
Net Recoverable assistance	repayment	sps_loa	Average amount of recoverable assistance received by clients in SPS
Net Recoverable assistance	repayment	slh_loa	Average amount of recoverable assistance received by clients in SLP-HCD
Net Recoverable assistance	repayment	a_loa	Average amount of recoverable assistance received by clients in EB, SLP-Carer and OB
Net Recoverable assistance	repayment	z_loa	Average amount of recoverable assistance received by clients in SUP
Net Recoverable assistance	repayment	nob_loa	Average amount of recoverable assistance received by clients not supported by benefit
Winter energy payment	Tier 2	wep_pmt	Average amount of winter energy payment received for those supported by a main benefit

Notes: refer to Section 9.5 for further details of the net recoverable assistance payments

AS payments and emergency housing payments are included in this section for completeness but refer to Section 10 for details of these models and their application.

Table 9.3 – Accommodation supplement and emergency housing payment models

Payment	Type	Model ID	Description
Accommodation supplement	Tier 2	hou_as	Average Accommodation Supplement payment for working age clients in PH for all benefit types
Accommodation supplement	Tier 2	acc_pmt	Average Accommodation Supplement payment for working age clients in the AS housing state for all benefit types
Emergency housing	Tier 3	hou_emh	Average emergency housing payment for a client who had at least one emergency housing payment in the quarter.

Some detailed comments on the payment models follow:

- Payments are allocated by client quarter, or proportionally if payment spells span multiple quarters. Further, most payments are scaled to March 2023 benefit levels, using an inflation index applied to benefit payments over the past 23 years. We have used past increases in DPB/SPS payment levels to infer these increases, which were previously CPI based and are now AWE based, with some additional ad hoc increases. The simulation model has an adjustment to adjust payments to March 2023 benefit levels in cases where the model was not refit for the 2023 modelling round and is therefore scaled to levels as at dates prior to March 2023.
- All models were Poisson with a log link. The choice of distribution was found to have a very minor effect on predictions in the payment models.
- As implied above, some payment models are 'shared' across benefit states– for example, the disability allowance for clients on JS-WR, EB, SLP-Carer and OB all use the 'a_da' payment model. This sharing is done when the individual models are share similarities to improve the efficiency of modelling. In these cases, the current benefit state is also used as a predictor to ensure that any differences between states are still modelled.
- It is possible to receive more than one Tier 1 benefit in a quarter. We have dealt with this by reallocating all Tier 1 payments to the current state; for example, if someone is allocated to JS-WR in a quarter but they receive both JS-WR and JS-HCD, all payments are summed and treated as JS-WR. The overall impact of this allocation is very small, since:
  - The amounts involved are generally small compared to a full quarter's benefit
  - The allocations largely offset each other (e.g. for every client with a JS-HCD payment allocated to JS-WR there is another with a JS-WR payment allocated to JS-HCD)
  - The average number of quarters before transitions is high enough that such a reallocation occurs in a relatively small proportion of quarters.
- NOB requires payment models for CCS, HS and EI because clients only in receipt of these benefits are assigned to the NOB state.
- There is an important point to note regarding the non-main payment models (that is, every model except Tier 1 in Table 9.2 and Accommodation Supplement in Table 9.3). These payments represent an average value across people in each benefit state; to take an example, the DA model for those in the JS-WR state estimates the average DA paid to clients receiving JS-WR, conditional on all their attributes like age, gender etc. However, some JS-WR clients receive DA and some do not, so at an individual level these payment models are misleading since the actual DA payments will usually be much higher (if the client receives DA) or much lower (if they do not). These payment levels are appropriate for the aggregate and segment level results but must be interpreted carefully when

inspected at an individual level. Distinguishing between the cases of receipt of supplementary payments at an individual level is beyond the scope of this projection.

- Prior to the introduction of the combined benefit/public housing projection, AS was estimated in a similar manner to the other benefit types, with AS payment models for each benefit type, averaged over all recipients of that benefit. Now we explicitly model the receipt of AS as a housing state. This enables more accurate individual-level estimates of AS support. Although AS receipt is not permitted while in a public house, it is possible to receive AS before or after being in a public house within a quarter – hence the need to have an AS model for both the PH and AS housing states.
- Note that the LOA model refers to recoverable assistance payments made to clients. These are later partly offset by recoveries of recoverable assistance – see Section 9.5 for further details.

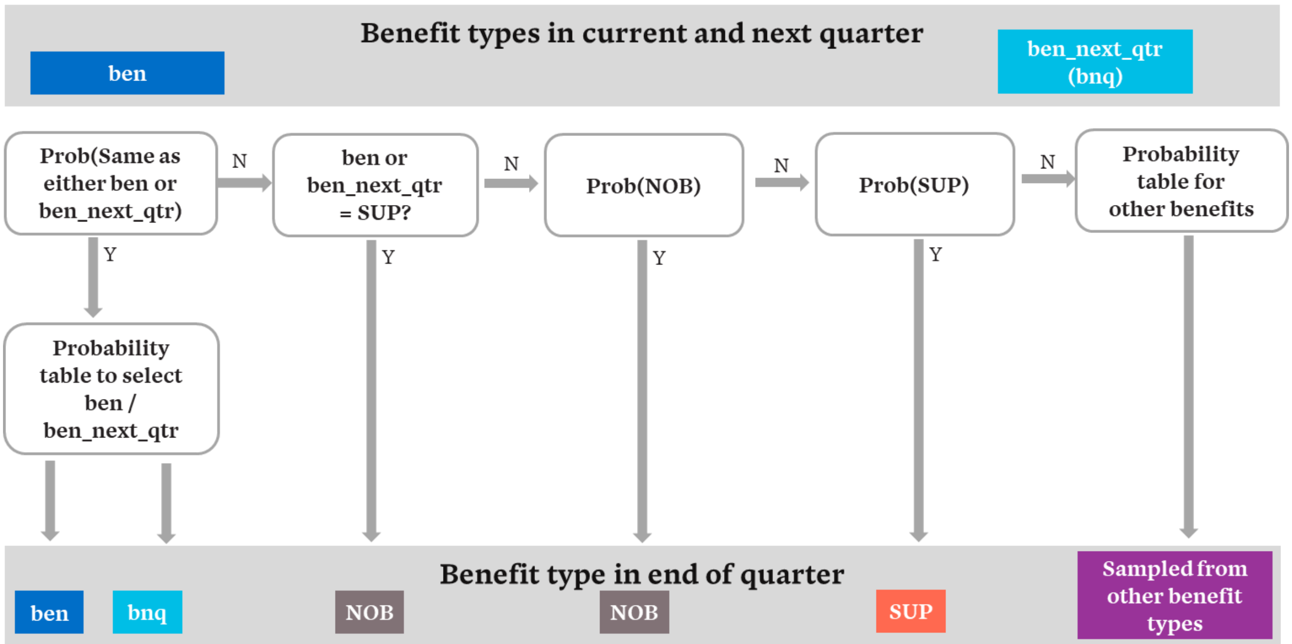
9.4 Welfare overlay models

9.4.1 Overview

As previously noted, the benefit state used in the projection is taken to be the benefit state during the quarter and not the benefit state at the end of the quarter. While this is valid for projection purposes, the end of quarter benefit state is important for reporting purposes. For example, knowing the benefit at the end of a quarter enables a count of the number of people receiving benefits. Although the benefit states during and at the quarter end often line up, this is not always the case. For example, end of quarter numbers can be between 70-99% of over the quarter numbers, e.g. 70% for new JS-WR clients and 99% for long-term SLP-HCD clients.

A set of chained models is used to predict the end of quarter benefit state given the states in the current and next quarter. These models are a mixture of probability GLM models and probability tables and are listed in Table 9.4 while the flowchart in Figure 9.4 sets out the order of application.

Figure 9.4 – Overlay models for end-of-quarter benefit



As well as models to determine the benefit at the end of the quarter, we also model the proportion of a quarter someone has spent supported by a main benefit. This is modelled using a one-inflated beta distribution – effectively meaning that different parameters are used for the probability that someone spends the entire quarter on a main benefit.

9.4.2 Detailed list of all models used

Table 9.4 – Welfare overlay models

Model	Type	Model ID	Description
ben_end is either ben or ben_next_qtr	Binomial GLM	eq_besame	Probability that end quarter benefit is either the benefit state in the current quarter or that in the next quarter.
Select ben or ben_next_qtr	Probability table	eq_samet	Probability table used to set ben_end equal to ben or ben_next_qtr conditional on it being one of these.
ben_end is NOB	Binomial GLM	eq_benob	Probability that ben_end is NOB conditional on it not being ben or ben_next_qtr and neither ben nor ben_next_qtr are SUP.
ben_end is SUP	Binomial GLM	eq_besup	Probability that ben_end is SUP conditional on it not being ben or ben_next_qtr and not NOB.
ben_end otherwise	Probability table	eq_othT	Probability table used to determine ben_end given it is none of the above.
Continuous duration resets	Binomial GLM	eq_reset	Probability that there is a break in benefits of more than two weeks during a quarter.
Proportion of quarter	One-inflated beta	ben_prop	Measures the proportion of the quarter that someone spent on a main benefit

9.5 Modelling recoverable assistance

Prior to moving to the IDI in 2018, we modelled net loans and expenses including recoverable assistance. Since 2018 we have retained the modelling of recoverable assistance, but not overpayments and operating assistance. These can be added back in if this information is useful.

We have used the term 'Recoverable Assistance' to include all types of benefits and assistance that are recoverable (excluding overpayments and fraud). Thus, Recoverable Assistance includes benefit advances and recoverable Special Needs Grants (SNGs), as well as a few minor related payments. In the provided data, the payments related to Recoverable Assistance are included under specific benefit codes. The costs associated with Recoverable Assistance relate to the non-recoverability of some assistance.

The following methodology has been used for Recoverable Assistance:

- Payments are estimated in the same fashion as other benefits and assistance
- Recoveries are estimated as a percentage of Recoverable Assistance payments.

This represents a relatively simple approach reflecting the fact that the non-recovered component of recoverable assistance represents a small part of overall benefit payments made.

9.6 Individual characteristic models

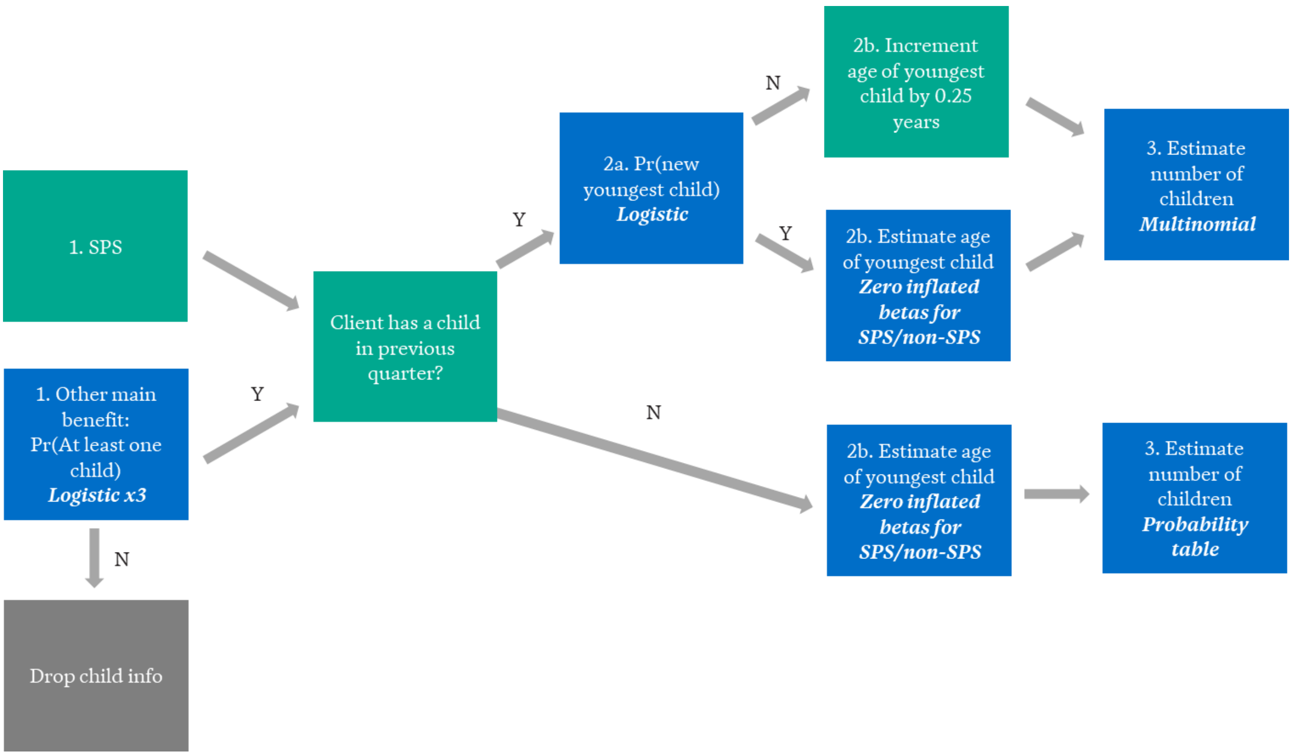
As well as the core transition and payment models for the welfare system we have a number of welfare system related variables that are modelled.

9.6.1 Children variables

We model child numbers for all clients receiving main benefits (i.e. also EB/SLP-Carer/SLP-HCD/JS-WR/JS-HCD). This is because children variables are important for all main benefits. Payments are often linked to number of children, while children information impacts transitions between and out of benefits. The variables modelled are the number of children up to a maximum of three and the age of the client's youngest child.

The process broadly follows three steps which are outlined in Figure 9.5 and detailed further below.

Figure 9.5 – Children variables flowchart



- **Determine if the client has at least one child.** This step was previously unnecessary as by definition all clients on SPS have at least one child. For clients not on SPS, we use a logistic GLMs to determine if they have at least one child. There are separate GLMs for different groups of clients:

  - Those who have newly entered main benefit (excluding SPS)
  - Those who remain on any main benefit (excluding SPS) and previously had at least one child
  - Those who remain on any main benefit (excluding SPS) and previously had no children.
- **For those with at least one child, determine the age of the youngest child.** The process for this depends on whether the client had any children in the previous quarter.
  - For those who were recorded as having at least one child in the previous quarter – a logistic GLM is used to determine if there has been a change in the age of youngest child other than expected (i.e. anything other than an increase of 0.25 years). If there is a change in the age of the youngest child, the value for the new age is sampled from one of two zero inflated beta models. Clients on SPS use a model that allows the age of youngest child to go up to 14 as this is the maximum child age of eligibility for SPS. The remaining main benefit types sample from a model that allows the age of the child to go up to 17.75. If there is no change in youngest child, the age of youngest child increases by 0.25 years.
  - For clients who did not have a child in the previous quarter – separate zero inflated beta models are used to determine the age of their youngest child. Again, clients on SPS use a separate model that allows children up to the age of 14 while all other clients share a model that allows children up to the age of 17.75.
- **For those with at least one child, determine the number of children.** There are three methods for this depending on the client's situation:
  - For those who have entered any main benefit – sample from a probability table
  - For those who were supported by a main benefit (excluding SPS) in the previous quarter and didn't have any children – sample from a separate probability table.
  - For those who had at least one child recorded in the previous quarter – use a multinomial to predict the new number of children. This is most likely the same number as the previous quarter. The age of youngest child and whether there has been a change in youngest child are important predictors in this model.

For clients who have left main benefits or are on a benefit but are determined to not have a child, all child information is dropped. Therefore, if they are to enter benefit again a new number of children and age of the youngest child will be simulated.

9.6.2 Partner flag

MSD captures partner information for clients receiving EB/SLP-HCD/JS-HCD/JS-WR. We model this partner information for people receiving main benefits. The variable is updated as follows:

- Moving into any of EB/SLP-HCD/JS-HCD/JS-WR from one of the other benefits: a binomial GLM estimates the probability that the client has a partner
- Remaining in any of EB/SLP-HCD/JS-HCD/JS-WR: a binomial GLM gives the probability that the partner flag switches (i.e. if the client has a partner, do they switch to having no partner and vice-versa)
- Leaving EB/SLP-HCD/JS-HCD/JS-WR and moving into one of the other benefits: partner information is dropped. This includes those who have retired.

9.6.3 Incapacity information

For clients receiving JS-HCD or SLP-HCD we keep track of information about their incapacity. These variables are:

- The type of incapacity

- The number of incapacities
- Whether the incapacity relates to their partner
- Whether they have a psychological incapacity (primary or secondary)
- Reassessment frequency (if receiving SLP-HCD only).

The incapacity variables are updated as follows:

- If a client enters JS-HCD or SLP-HCD then the above variables are sampled from 5 different sets of models for each variable. The majority of these models are probability tables rather than GLMs.
- There are different models (all probability tables, unless otherwise specified) for each of the situations:
  - entry into JS-HCD except from SLP-HCD
  - entry into SLP-HCD except from JS-HCD
  - switching from JS-HCD to SLP-HCD
  - switching from SLP-HCD to JS-HCD
  - remaining in SLP-HCD or remaining in JS-HCD.
- If a client remains in JS-HCD or SLP-HCD a binomial GLM gives the probability that the client changed primary incapacity type. If so, then the remaining models to simulate the new incapacity variables are probability tables.

On leaving JS-HCD / SLP-HCD: incapacity variables are forgotten. This applies to retirees as well.

9.6.4 Detailed list of all models used

Table 9.5 – Welfare individual characteristic models

Model	Type	Model ID	Description
Children, benefit entry	Binomial GLM	chd_bh_enty	Probability of having at least one child for those entering benefit.
Children, previously no children	Binomial GLM	chd_bh_new	Probability of having at least one child if supported by benefits and had no children in the previous quarter.
Children, previously children	Binomial GLM	chd_bh_none	Probability of having a child if supported by benefits and had a child in the previous quarter.
Number children on entry	Probability table	chd_bh_enum	Number of children for those entering benefit.
Number children, previously no children	Probability table	chd_bh_nnum	Number of children for those who didn't have children in the previous quarter.
Number children, previously children	Multinomial	chd_bh_num	Number children for those who remain supported by benefits and had children in the previous quarter.

Model	Type	Model ID	Description
New youngest child	Binomial GLM	chd_bh_age	Probability changing youngest child.
Age youngest child, entry SPS	Beta	sps_aye	Age of youngest child on entry to SPS
Age youngest child, continuing SPS	Beta	sps_aye	Age of youngest child for those remaining in SPS
Age youngest child, entry	Beta	chd_bh_aye	Age of youngest child on entry to benefit (excludes SPS)
Age youngest child, continuing	Beta	chd_bh_ayr	Age of youngest child for who remain supported by benefit (excludes SPS)
Partner, entry	Binomial GLM	a_prt1	Probability of having a partner for those entering EB, JS-WR, JS-HCD, SLP-HCD.
Partner, no previous	Binomial GLM	a_prt2	Probability of having a partner for those **without** a partner in the previous quarter in EB, JS-WR, JS-HCD, SLP-HCD.
Partner, previous	Binomial GLM	a_prt3	Probability of having a partner for those **with** a partner in the previous quarter in EB, JS-WR, JS-HCD, SLP-HCD.
Incap, change	Binomial GLM	a_incp	Probability of changing primary incapacity type for those remaining in JS-HCD and SLP-HCD.
Incap, to JS-HCD not from SLP-HCD	Probability tables	semi_incap1_new	Probability of incapacity variables changing for those entering **JS**-HCD who were **not** in **SLP**-HCD in the previous quarter.
Incap, to SLP-HCD not from JS-HCD	Probability tables	semi_incap2_new	Probability of incapacity variables changing for those entering **SLP**-HCD who were **not** in **JS**-HCD in the previous quarter.
Incap, to JS-HCD from SLP-HCD	Probability tables	semi_incap3_new	Probability of incapacity variables changing for those entering **JS**-HCD who were in **SLP**-HCD in the previous quarter.

Model	Type	Model ID	Description
Incap, to SLP-HCD from JS-HCD	Probability tables	semi_incap4_new	Probability of incapacity variables changing for those entering **SLP**-HCD who were in **JS**-HCD in the previous quarter.
Incap, remain	Probability tables	semi_incap5_new	Probability of incapacity variables changing for those remaining in **SLP**-HCD and **JS**-HCD.

9.7 Changes since the previous report

None.

10 Public Housing models

This section describes the models specific to modelling public housing-related outcomes. It covers:

- The housing state construct
- The models that project transitions between different states
- Payment models – the modelling of the amount of IRRS, AS, EMH and TAS payments.

A full list of the models is included.

10.1 Introduction

The public housing component of the projection models:

- Individuals in Public Housing (PH)
- Individuals receiving Accommodation Supplement (AS)
- Movements of individuals to and from the housing register
- Individuals receiving emergency housing (EMH).

As well as recording individual information (housing state, role in household, TLA, register status etc.), the projection retains information about those in public housing at the start of the projection and projects their evolution through time, including exits from public housing.

Much of the housing component operates under a similar framework to the welfare projection: housing related variables are updated, housing transitions are forecast, payments are calculated for each individual. However, there are two complicating factors in the housing part of the projection that require special handling:

- The aforementioned tracking of households in existence at the projection date
- Supply and demand mechanics of the public housing system.

The rest of this section provides more details around the various elements of the housing component. Deterministic semi-dynamic variables, such as duration since being in a public house or duration since being on the register, update in a logical fashion. For example, in the latter case, the duration since being on the register resets to zero if the individual is on the housing register, otherwise it increments by one each quarter.

10.2 Transition models

Housing transitions are more complicated than welfare transitions. In particular, individual transitions are not independent due to housing supply and demand constraints. To capture housing movements, including register applications and public housing allocations, there are three phases in the housing transition model:

- Phase 1: forecast transitions to next quarter's housing state ignoring any register movements.
- Phase 2: forecast register applications in the current quarter and then allocate houses based on available supply.
- Phase 3: recalculate transitions for any individual allocated to a public house based on their new circumstances. Also estimate IRRS for those assigned to a new public house.

A further complication in the housing model is that transitions for individuals starting in a public house at the projection date are considered jointly to enable us to better model the evolution of households over time.

The following subsections discuss each of these points in more detail.

10.2.1 Phase 1 – transitions ignoring any register movements

Housing transitions are calculated in three phases. In the first phase, the housing state in the next quarter is calculated assuming there are no register movements (applications to the register, allocation of a public house).

Figure 10.1 and Figure 10.2 set out the process flows for this step.

Figure 10.1 – Housing transitions from public housing - Phase 1

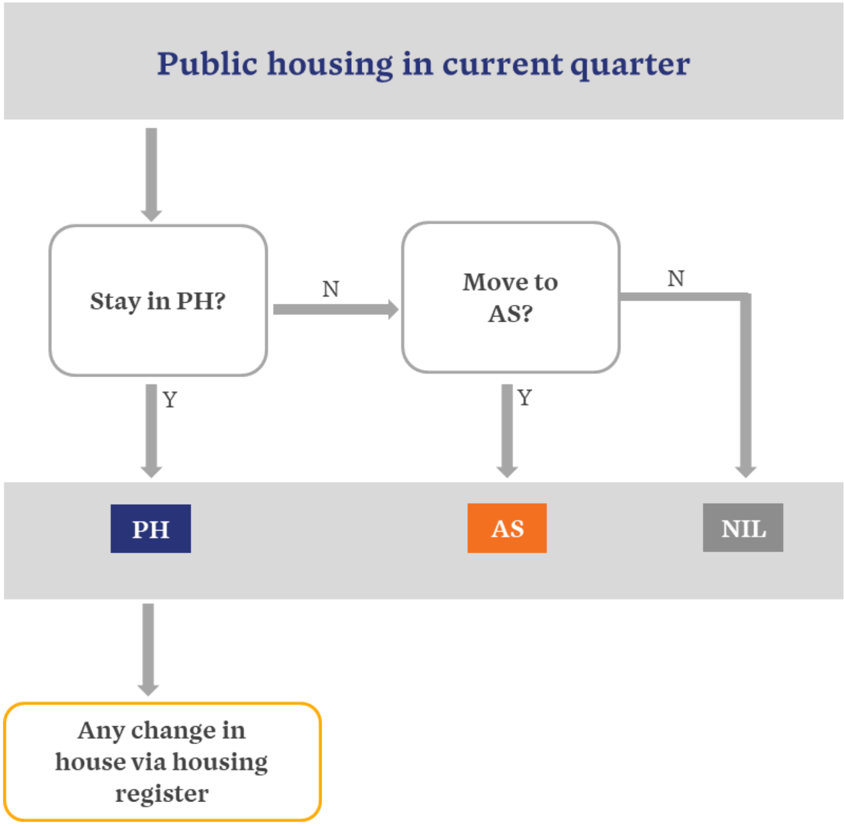
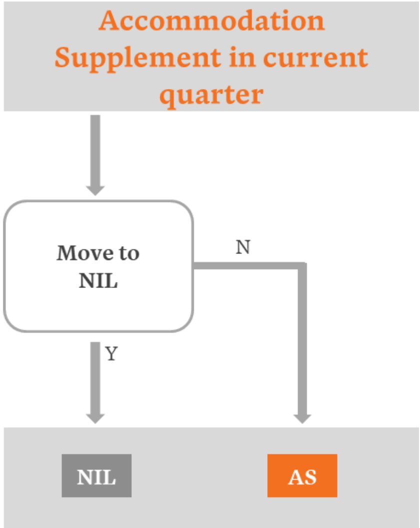
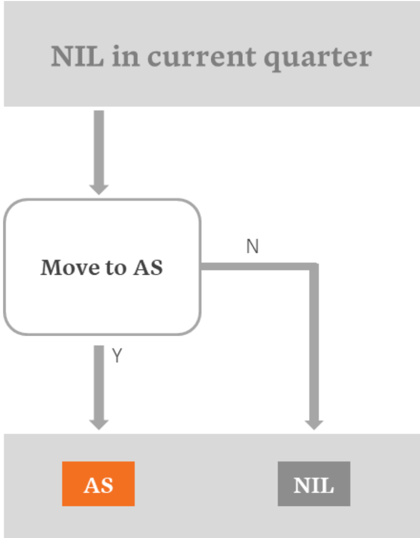


Figure 10.2 – Housing transitions from AS and Nil - Phase 1





In more detail, the steps are:

- Assuming the individual remains in the projection after mortality transitions, the next step is to calculate their housing state next quarter. The key thing to note is that individuals in AS or NIL (i.e. not in PH and not currently receiving AS) cannot transition to PH during this phase of the projection as transitions to PH can only happen via the housing register and house allocation processes.
- The housing state in the next quarter calculated here is the final result for those who do not have an active register application and who do not move onto the housing register in the next phase (for all those who end up on the register, the housing state may change if they are allocated a public house). This means that anyone in PH who is forecast to be in PH in the next quarter and who is not on (or does not move onto) the register remains in their current house.
- Those currently in PH who are forecast to move to AS or NIL in the next quarter must vacate their house. Given that an individual is classed as PH if they spend any time in a public housing during a quarter, this means that the individual must leave the house in the current quarter and, where a primary signatory has left the house, we assume that the house has been vacated by all occupants and place the house into the pool of houses available for reallocation.

10.2.2 Phase 2 – Register applications and housing allocation

10.2.2.1 Register

The next step is to create the housing register for the quarter. This consists of the following people:

- Those who remained on the register from the previous quarter
- Those who make new register applications in the current quarter. All individuals in the projection that are not currently on the register may apply. Note that different models are used for transfer register applicants (i.e. those currently in a public house) and other applicants (those not currently in a public house).

New register applicants require various register and priority characteristics. Their System Allocation Score (SAS) priority level (A/B) is determined using a GLM. A probability table is used to determine the TLA code (board in Auckland) where they have applied for a public house while other characteristics (signatory type, size of house, other SAS variables) are sampled from a representative population by age, region and SAS priority level.

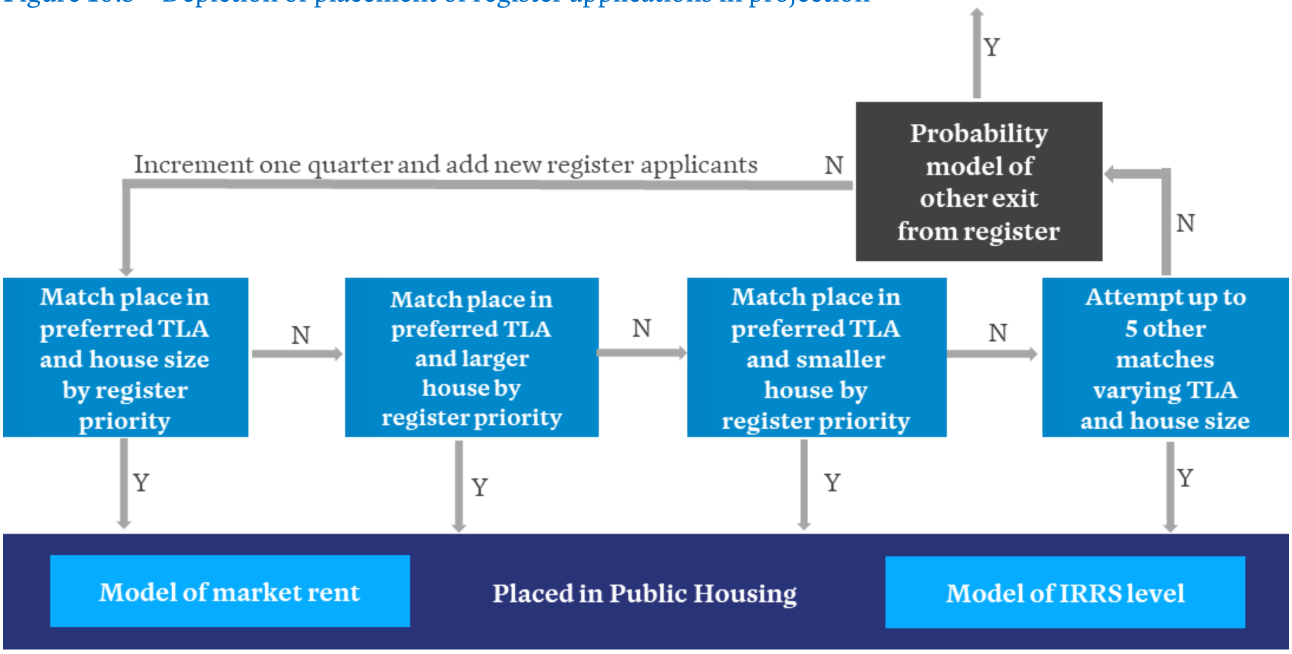
For each individual on the register, a model is used to determine the relative likelihood that clients move from the register to public housing. This likelihood model is data-based, and primarily depends on SAS priority and related underlying need scores. The register is then sorted using a randomly perturbed version of this likelihood to produce the priority ordering used for housing allocation.

10.2.2.2 Housing allocation

Houses are then allocated based on the priority ordering of the register and subject to housing availability. The process is represented graphically in Figure 10.3 and is as follows:

- First, match based on the preferred TLA and house size, taking the priority ordering of the register into account.
- Next, for those remaining on the register, allocate houses based on preferred TLA but larger houses (one bedroom larger) than requested.
- At the third attempt use the preferred TLA and a smaller house (one bedroom less).
- Make up to five other attempts to allocate houses by varying the TLA (board) and house size. The number of attempts varies based on house type requested and SAS priority, with more attempts made to house priority A applicants.

Figure 10.3 – Depiction of placement of register applications in projection



Note that this housing allocation process considers tranches of people at the same time, rather than on an individual basis. The effect of this is to prioritise allocating houses by first preference to as many people as possible and, in practice, the majority of allocations are houses meeting an individual's first choice.

10.2.3 Phase 3 – Consequences of housing allocation

In Phase 3, the projection deals with the consequences of house allocation for those who have successfully received a house. First, using models of market rent and IRRS levels, the IRRS payment is determined for each signatory individual. Secondly, housing transitions are recalculated based on the newly allocated house.

In particular, note that the housing allocation process means that the individual's housing state may be altered for the current quarter as well as the next quarter. This is due to the hierarchical nature of the definition of the PH state – any spell in PH, no matter how small, during a quarter will result in the individual being listed in PH. This is in contrast to all other transitions (welfare benefit types and AS), where the state in the current quarter is fixed and the state in the next quarter is the unknown quantity. There is justification for this approach as it deals with the situation where an individual not currently in PH applies to the register and is allocated a house in the same quarter.

For those on the register who do not get allocated a house, a model is used to decide if they remain on the register. If they do, then further attempts to house them will be made in the next quarter.

10.2.4 Treatment of households

Households evolve over time; children leave home, singles become partnered and couples can split. The grouping of individuals into households in future years is difficult. Further, the data available are scant; household evolution while in public housing is available, but there is little data for what happens after exiting public housing. For the purposes of this report, we have simplified the treatment of households for tractability. A similar simplification was used in previous reports.

- Existing households in public housing at the projection date are modelled as a group, the movements of one householder will be closely related to the movements of the primary householder. We refer to these as 'real' households.
- Future households are notional; we model people as individuals and assign them to notional households ('Person A is the partner in a household of size four'), but we do not formally link these individuals as a household. Note that individuals consist of both primary and non-primary household members to ensure a full coverage of individuals in a public house in the projection.
- Real households are dissolved when they exit, as there is limited ability to track household status. Therefore, all householders become notional. This means for a couple who exit public housing, the future housing state of one is not affected by the other.

10.2.5 Emergency housing

Emergency housing is not strictly part of the transition models as it is possible to receive emergency housing in the same quarter as being in Public Housing, receiving Accommodation Supplement or neither. Rather it is an additional flag that is informed by the current housing state of the client.

The flag has been fitted using a GLM that measures the probability of an individual receiving an emergency housing payment. There is limited data available to fit the model as:

- Emergency housing payments have only been in existence since 1 July 2016
- Relatively few clients receive emergency housing in a quarter.

We also include an individual's history of emergency housing receipt over the prior year as a predictor to improve the performance of the emergency housing flag model.

10.2.6 Detailed list of all models used

The tables below show the models that describe the transition behaviour in the public housing system. All probability models are binomial GLMs with the standard logistic link.

Table 10.1 – List of housing transition models used in projection

Model	Housing state	Type	Model ID	Description
PH exit	PH	Logistic	hou_tra	Probability that a client in a public house and aged <65 remains in a public house the following quarter
PH exit	PH	Logistic	hou_trap	Probability that a client in a public house and aged >64.75 remains in a public house the following quarter
PH exit	PH	Logistic	hou_acc	Probability that a householder aged <65 and in a public house and exits the public house receives AS the following quarter

Model	Housing state	Type	Model ID	Description
PH exit	PH	Logistic	hou_accp	Probability that a householder aged >64.75 and in a public house and exits the public house receives AS the following quarter
PH exit	PH	Logistic	hou_sec	Probability that a non-primary householder remains in a public house given the primary householder exits
PH exit	PH	Logistic	hou_sec2	Probability that a non-primary householder remains in a public house given the primary householder remains
AS exit	AS	Logistic	acc_nil	Probability that an AS client aged <65 does not receive AS in the next quarter, given the client does not move into a public house
AS exit	AS	Logistic	acc_nilp	Probability that an AS client aged >64.75 does not receive AS in the next quarter, given the client does not move into a public house
AS entry	NIL	Logistic	nil_acc	Probability a client aged <65 who is not 'Not on benefit' (NOB) receives AS in the next quarter, given they do not move into a public house
AS entry	NIL	Logistic	nil_accp	Probability a client aged >64.75 who is not 'Not on benefit' (NOB) receives AS in the next quarter, given they do not move into a public house
Exit register	PH, AS or NIL	Logistic	reg_hou	Probability a client moves from the register to a public house
Exit register	PH, AS or NIL	Logistic	reg_oth	Probability a client exits the register not to a public house
Enter register	PH	Logistic	tran1	Probability a client in a public house makes a register application in the quarter
Enter register	AS or NIL	Logistic	reg1	Probability a client not in a public house makes a register application in the quarter
Register priority	PH	Logistic	tranAB	Probability that someone on the transfer register is SAS_priority A

Model	Housing state	Type	Model ID	Description
Register priority	AS or NIL	Logistic	reg_A	Probability that someone on the new register is SAS_priority A
Emergency housing	PH, AS or NIL	Logistic	hou_ehfl	Probability that someone has at least one emergency housing payment in the quarter
Preferred register location	PH	Table	semi_reg_tla_transfer	Table for selecting TLA code for register application for individuals already in PH
Preferred register location	AS or NIL	Table	semi_reg_tla_new	Table for selecting TLA code for register application for individuals not already in PH
Preferred register location	PH, AS or NIL	Table	semi_reg_tla_all	Fallback table for selecting TLA code for register application for individuals not matched in the semi_reg_tla_transfer or semi_reg_tla_new tables

10.3 Payments

Once the housing allocation process has occurred, payments relating to public housing and accommodation support may then be calculated based on the housing state in the current quarter and the current TLA code. There are four housing payment types:

- IRRS (PH only)
- Accommodation Supplement (AS)
- Temporary Additional Support (TAS)
- Emergency Housing (EMH).

10.3.1 IRRS

This is applicable only to signatory householders in PH. For clients who enter a new house in the current quarter, a model is used to first set the market rent, followed by a model to set the IRRS level relative to this.

For those that remain in a house we have a series of models for IRRS updating each quarter:

- Probability that IRRS level moves from zero to nonzero, or vice versa
- If it toggles to nonzero, a probability table for expected IRRS level (as a fraction of market rent)
- If IRRS remains nonzero, a probability model for whether the new IRRS equals the default update. If not, apply a probability table for the new IRRS level.

The default update is calculated as:

$$\begin{aligned} Default\ IRRS\ update \\ &= (Old\ rent \times (rental\ growth\ \ inflation - AWE\ inflation)) \\ &+ old\ IRRS \times AWE\ inflation \end{aligned}$$

IRRS payments are then calculated. They are allocated to individuals based on the number of signatories in the household – so a house with two signatories will see half the calculated IRRS amount attaching to each signatory individual.

The projection of IRRS payments is at an individual level, though IRRS costs are allocated evenly across signatories in the household. It is possible to obtain household-level lifetime cost estimates for households in existence at the projection date. Note that these costs include future housing support to current non-signatories. So, if a non-signatory leaves a current household and re-enters public housing as a signatory in a new household, this will contribute to the lifetime cost estimate of the initial household.

There are alternatives for attaching IRRS payments to household members. We believe our current approach is a reasonable basis for operational intervention, as the income for signatories is the main determinant of the level of IRRS support.

10.3.2 AS, TAS and EMH

AS, TAS and EMH are calculated using GLMs. Different models are used for AS and TAS for those of working age and older individuals.

10.3.3 Detailed list of models

The table below sets out the models used to calculate housing payments. Models are a mixture of GLMs, zero inflated beta and probability tables.

Table 10.2 – List of housing transition models used in projection

Model	Housing state	Type	Model ID	Description
IRRS per week (remain)	PH	Logistic	hou_irrs3	Probability that weekly IRRS toggles between zero and non-zero
IRRS per week (remain)	PH	Logistic	hou_irrs4	Probability that those with positive IRRS do not change in the current quarter
IRRS per week (remain)	PH	Table	semi_irrs1	Ratio of weekly IRRS to weekly market rent given that IRRS is forecast to change from zero for individuals remaining in a public house
IRRS per week (remain)	PH	Table	semi_irrs2	Ratio of weekly IRRS to weekly market rent given that IRRS is forecast to change from a non-zero level for individuals remaining in a public house
IRRS per week (new)	PH	Linear	market1	Mean of individuals weekly market rent for an individual newly allocated a public house
IRRS per week (new)	PH	Zero inflated beta	market2	Ratio of weekly IRRS to weekly market rent for an individual newly allocated a public house

Model	Housing state	Type	Model ID	Description
IRRS per qtr (entry)	PH	Beta	hou_irrs1	Estimate of IRRS for the quarter for people new to public housing
IRRS per qtr (entry)	PH	Beta	hou_irrs1p	Same as hou_irrs1, but for individuals aged 65 or more
IRRS per qtr (existing)	PH	Beta	hou_irrs2	Estimate of IRRS for the quarter for people continuing in public housing
IRRS per qtr (existing)	PH	Beta	hou_irrs2p	Same as hou_irrs2, but for those aged 65 or more
AS payments	PH	Poisson GLM	hou_as	AS payments while in PH state for those aged <65
AS payments	PH	Poisson GLM	hou_asp	AS payments while in PH state for those aged 65 or more
AS payments	AS	Poisson GLM	acc_has	AS payments while in AS state for people aged <65
AS payments	AS	Poisson GLM	acc_hasp	AS payments while in AS state for people aged 65 or more
TAS payments	PH	Poisson GLM	hou_tas	TAS payments while in PH state
TAS payments	AS	Poisson GLM	acc_tas	TAS payments while in AS state for those aged <65
TAS payments	AS	Poisson GLM	acc_tasp	TAS payments while in AS state for those aged 65 or more
TAS payments	NIL	Poisson GLM	nil_tas	TAS payments while in NIL state
EMH	PH, AS or NIL	Poisson GLM	hou_emh	Average emergency housing payment for all clients who received an emergency housing payment

10.4 Changes since the previous report

None.

11 Income models

This section describes the models specific to modelling incomes and related amounts. It covers:

- Personal income
- Working for Families tax credits
- NZ superannuation payments

A full list of the models is included.

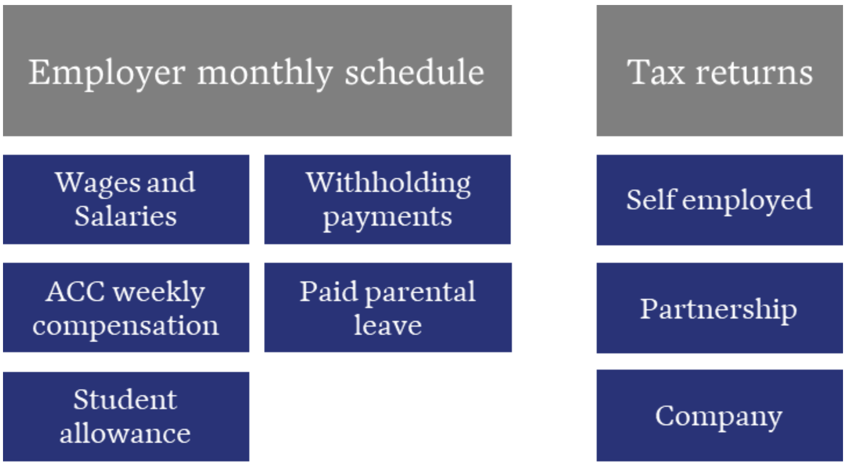
11.1 Personal income

Previously personal income was included as an overlay, meaning that it did not impact any other models. Since the 2019 model, personal income is included as a predictor. In fact, it has led to a significant improvement to the quality of fit in most models.

11.1.1 Make-up of personal income

Personal income consists of several different types of taxable income. The data for these all come from two tables by Inland Revenue.

Figure 11.1 – Make-up of personal income



Tax return data is only recorded on an annual basis. To get this on the quarterly basis of our model we spread it evenly throughout the year for periods when an individual has not received a benefit.

11.1.2 Working age population

The methodology for predicting personal income for those who are working-age (less than 65) is the same as that used in the previous report. Income is banded into four different amount groupings:

- Band 1: No income in the quarter (defined as less than $2)
- Band 2: Some income but less than half the minimum wage at 40 hours per week
- Band 3: Income of more than 50% but less than 100% of 40 hours per week at minimum wage
- Band 4: Income greater than 100% of minimum wage at 40 hours per week

The bands have been set with respect to the minimum wage based on the date of the data for the modelling fitting process. They are assumed, in our projection, to increase in line with average weekly earnings (AWE).

For the 2023 model this was data at 31 March 2023. The minimum wage at this time was $21.20 per hour. The current value date for starting bands and income values in the projection model is 31 March 2023.

Table 11.1 – Income bands

Band	Min	Max
Band 1		$2
Band 2	$2	$5,512
Band 3	$5,512	$11,024
Band 4	$11,024	

Note that the bands are strictly less than the maximum amount

In fitting the models all income amounts were inflated using AWE to 31 March 2023. Bands were derived using these inflated values.

11.1.2.1 Modelling

Once the income band for an individual is determined amounts can then be modelled. Each band uses a separate amount model. The exception is income band 1, where income is assumed to be equal to zero.

A transition modelling approach has been used to project which income band an individual is in. The methodology is shown in Figure 11.2 and Figure 11.3. Industry modelling also fits within the process after it is estimated whether someone earned positive income in the quarter and is explained further in Section 12.6.1.

Figure 11.2 – Personal income transition from income band 1

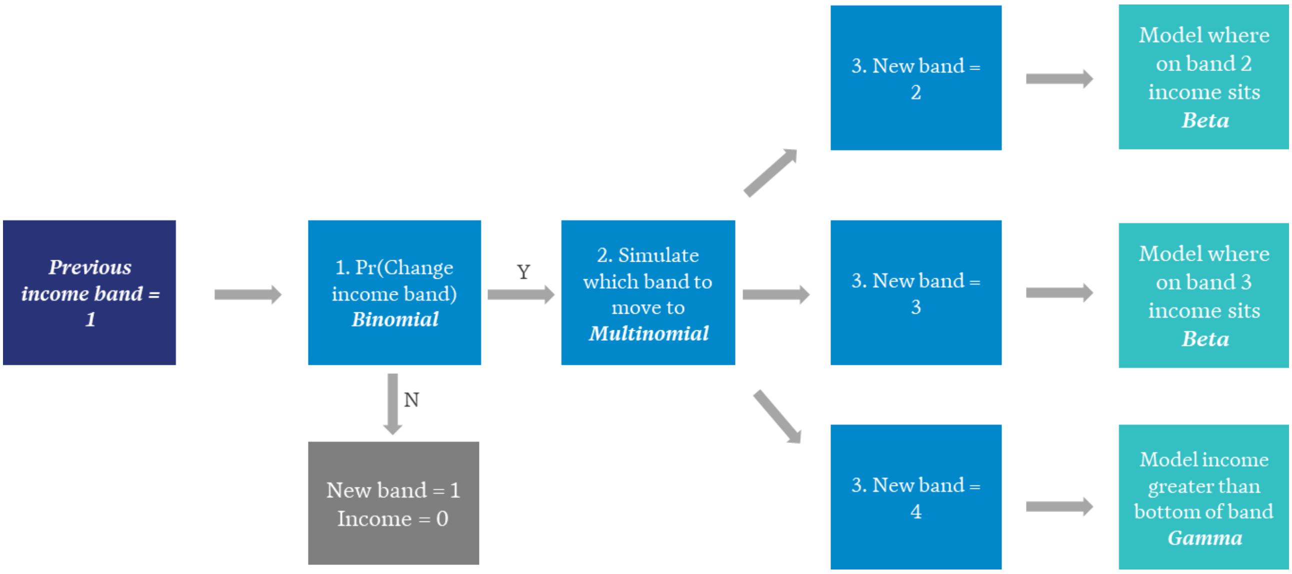
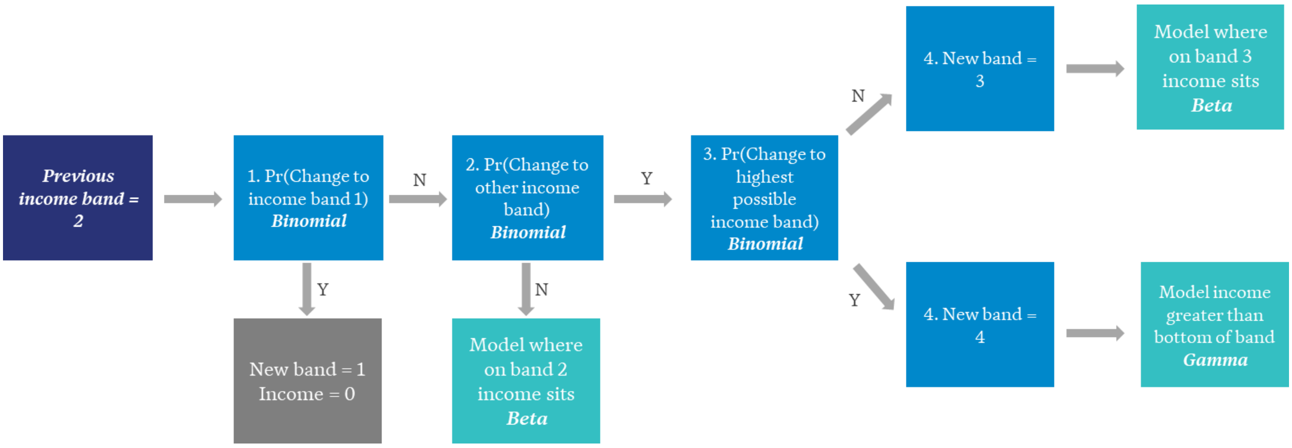


Figure 11.3 – Personal income transition from income band 2



The structure for income bands 3 and 4 is the same as income band 2.

11.1.3 Pensioners

Those aged 65 or older have very different income dynamics than those who are working-age. They are much more likely to transition to zero personal income, rather than moving between bands. Because of this we have kept the same approach for this segment since the 2018 modelling.

This approach is to have two models:

- First, model whether personal income is zero
- Given positive income, estimate the income amount.

11.2 Working for Families tax credits

Working for Families tax credits include:

- Family tax credit
- In-work tax credit
- Minimum family tax credit
- Parental tax credit.

Information inside the IDI is limited on the breakdown of paid amounts between the different types of tax credits. Due to this we have modelled WFF tax credits in aggregate rather than broken down by the type of tax credits. The information is only recorded on a tax return basis (March years) and we have spread this evenly across the four quarters of the year. Therefore, any consideration of WFF tax credits is best done on an annual basis and any seasonal effects will not be visible.

WFF tax credit eligibility depends primarily on:

- Household income (including benefits) in the year to March
- Family make up.

Neither of these items are explicitly modelled, so the WFF tax credit model uses proxies for these. For household income the main proxy is personal income in the previous year, while for family make-up there are a variety of proxies that may be used such as age, gender, ethnicity and benefit type received.

The partner and child data contained in the WFF tables is not suitable to be used in its current format without significant additional cleaning work. The state of the data prevented it being possible to allocate the tax credits to different members of the household. Due to this, when modelling WFF tax credits we have allocated the full tax credit to the primary individual listed in the household. If the partner and child data were able to be used, we would have looked at:

- Splitting tax credits by person by the proportion of total income of each household member
- Splitting tax credits by quarter by the proportion of total income of each person in each quarter.

Doing both of the above would result in a much stronger relationship between income and WFF tax credits and a better model.

11.3 NZ Superannuation

NZ Super has been modelled using a simple average payment model. This reflects that:

- A very high proportion of the resident over 65-year-old population receive it
- Most people receive one of four different amounts (single living alone, single living with someone, partner with partner over 65, partner with partner under 65).

Given that partnered status and living arrangements are not characteristics we can identify in the data for the non-working-age population, the model uses age, gender, ethnicity and region as predictor variables. These act as proxies, predominantly for partnered status e.g. the older a person is the less likely they are to have a partner. Ethnicity is included, not because ethnicity itself dictates NZ Super amounts, but because mortality rates differ by ethnicity (and hence reflect the likelihood of a partner being present). At younger ages, different ethnicities also tend to have different take-up rates. This relates to the criteria of living in New Zealand for a minimum number of years.

NZ Super payments are made fortnightly, generally on a Tuesday. This results in the quarterly payment being highly seasonal as it depends on the number of payments made in that quarter. We have not explicitly allowed for the number of payments to be made in the quarter as this is immaterial to the overall model relative to the additional work that would be required. Instead, we have allowed for the average seasonality by quarter.

11.4 Detailed list of all models used

Table 11.2 – List of income models used in projection

Model	Type	Model ID	Description
Income band 1	GLM	inc1_tra	Probability that a person previously in income band 1 moves into a new band this quarter
Income band 1	Multinomial	inc1_mul	For a person that moves out of income band 1 which income band do they move to
Income band 2	GLM	inc2_tra0	Probability that a person previously in income band 2 moves into income band 1 this quarter
Income band 2	GLM	inc2_tra	Probability that a person previously in income band 2 moves into income band 3 or 4 this quarter given they did not move to income band 1
Income band 2	GLM	inc2_inc4	For a person that moves out of income band 2 do they move to income band 4.
Income band 2	Beta	inc2_amt	Income amount for a person currently in income band 2

Model	Type	Model ID	Description
Income band 3	GLM	inc3_tra0	Probability that a person previously in income band 3 moves into income band 1 this quarter
Income band 3	GLM	inc3_tra	Probability that a person previously in income band 3 moves into income band 2 or 4 this quarter given they did not move to income band 1
Income band 3	GLM	inc3_inc4	For a person that moves out of income band 3 do they move to income band 4.
Income band 3	Beta	inc3_amt	Income amount for a person currently in income band 3
Income band 4	GLM	inc4_tra0	Probability that a person previously in income band 4 moves into income band 1 this quarter
Income band 4	GLM	inc4_tra	Probability that a person previously in income band 4 moves into income band 2 or 3 this quarter given they did not move to income band 1
Income band 4	GLM	inc4_inc3	For a person that moves out of income band 4 do they move to income band 3.
Income band 4	GLM	inc4_amt	Income amount for a person currently in income band 4
65 and over	GLM	pen_zinc	Probability of zero income in the quarter for people aged 65 or over
65 and over	GLM	pen_sinc	Average income in the quarter for people aged 65 or over, given they received positive income.
WFF tax credits	Poisson GLM	ftc_amt	Total amount of WFF tax credits in the quarter
NZ Super amount	Poisson GLM	pen_pen	Total amount of NZ Superannuation paid to those over 65 in the quarter.

11.5 Changes since the previous report

None.

12 Other models

One of the major benefits of modelling inside the IDI is that it provides access to linked data from many different sources. As our work inside the IDI has progressed, we have added more 'other models' outside the benefit and housing system. These are used as both predictors in other models, variables to summarise by and additional outputs that we can analyse.

12.1 Care and Protection and Youth Justice

The care and protection (CNP) for children and Youth Justice (YJU) variables that are used in the model align with the definitions Oranga Tamariki use in their 'Children's model'.

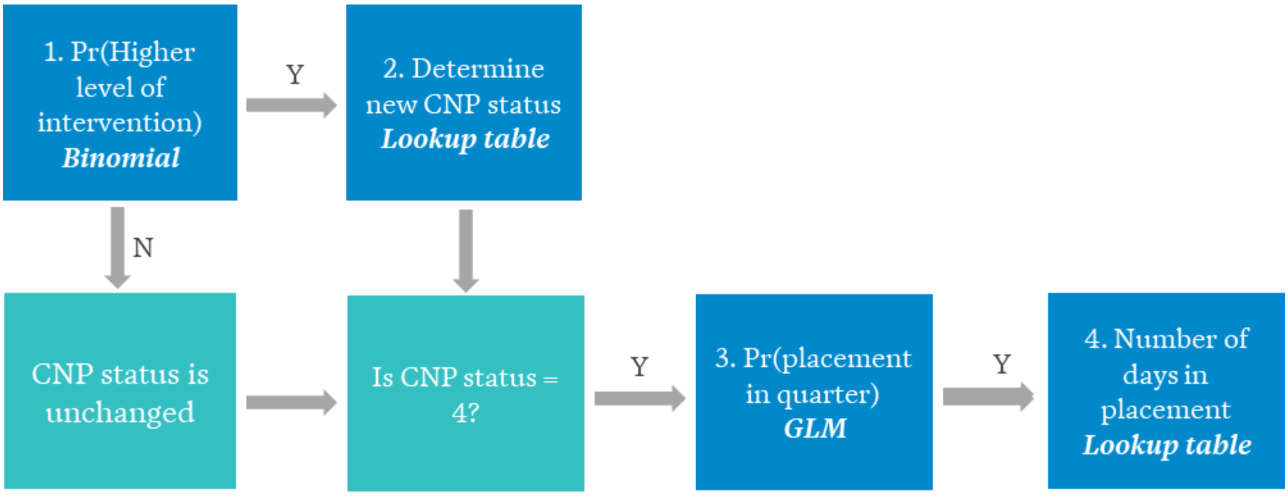
12.1.1 Methodology

The variables are updated for CNP for individuals up to age 18 as follows:

- A binomial GLM is run for the probability that the individual had a higher level of CNP intervention in the quarter then they ever had previously
  - If yes, then a lookup table is used to determine the new highest level of CNP intervention for the individual
- If the current highest level of CNP intervention is equal to 4 (placement) then another GLM is run for the probability they had a placement in that quarter
  - If yes, then a lookup table is used to determine the number of days the individual spent in a placement in the quarter

The diagram below shows the process as a flowchart. Variables are not retained after age 30.

Figure 12.1 – Care and protection flowchart



The process above is the same for YJU with separate models used.

12.1.2 Detailed list of all models used

Table 12.1 – List of care and protection and Youth Justice models used in projection

Model	Type	Model ID	Description
CNP status	GLM	cyf_cnp	Change in highest level of care and protection intervention
CNP status	Probability Table	cyf_cnp_lvl	New highest level of care and protection intervention
CNP days	GLM	cyf_srs	Statutory care and protection intervention in the quarter
CNP days	Probability table	cyf_cnp_days	Number of days for care and protection given statutory intervention in the quarter
YJU status	GLM	cyf_yju	Change in highest level of Youth Justice intervention
YJU status	Probability table	cyf_yju_lvl	New highest level of Youth Justice intervention
YJU days	GLM	cyf_yju_change_days	Probability of change in Youth Justice days
YJU days	Probability table	cyf_yju_days	Number of days in Youth Justice in quarter given a change in the number of days

12.2 Education

12.2.1 Secondary education

The process for updating secondary education is as follows while an individual is still at school:

- The probability that an individual leaves school during the quarter is estimated using a probability table. At age 21 the probability is 1 to ensure all simulated pathways leave school at that point.
- If the client leaves school during the quarter the NZQF level at exit is sampled from a probability table.
- If the client leaves school during the quarter, the probability of being stood down or suspended while at secondary school is estimated using a logistic GLM. The total number of days stood down or suspended is sampled from a Log-Normal distribution.

12.2.2 Tertiary education

Note that our definition of tertiary education includes any industry training as well.

12.2.2.1 Enrolled in study

One of the new variables we introduced in the previous report is whether an individual is in tertiary education. This is determined as follows:

- If not in tertiary education in the previous quarter: a logistic GLM models the probability of being in tertiary education in this quarter

- If in tertiary education in the previous quarter: a logistic GLM models the probability of not being in tertiary education in this quarter.

We also keep track of how long someone has been in study. Quarters in which individuals enter/exit study are key transition points for the benefit system as well for income band transitions.

This is modelled for all people who are in the working-age population which covers almost all people in tertiary education.

12.2.2.2 Level enrolled

We next determine the highest level of enrolment an individual has ever been in tertiary education. This is determined as follows:

- If someone is currently in study, then a logistic GLM models the probability that someone is enrolled in a new highest level. For people who have entered tertiary education for the first time then this is set to 1.
- Given that they are now studying at a higher level, a probability table is used to determine what that new highest level enrolled is.

Enrolled level is only estimated up to the age of 30 (inclusive).

12.2.2.3 Level attained

We also project what an individual's highest level of attainment is. This is determined as follows:

- If someone is currently in study, then a logistic GLM models the probability that someone attains a new highest level. For people who have entered tertiary education for the first time then this is set to 1.
- Given that they have now attained at a higher level, a probability table is used to determine what that new level of attainment is.

Attainment level is only estimated up to the age of 30 (inclusive).

12.2.3 Detailed list of all models used

Table 12.2 – List of education models used in projection

Model	Type	Model ID	Description
Secondary	Probability Table	edu_slve	Probability of leaving secondary education during the qtr
Secondary	Probability Table	edu_snqf	NZQF level at exit, given they leave secondary education during the qtr.
Secondary	GLM	edu_sdur	Probability of being stood down or suspended at secondary school given they leave secondary education during the qtr.
Secondary	Log Normal distribution	NA	Total number of days being stood down or suspended at secondary school given that they were stood down or suspended
Tertiary Education	GLM	edu_tnt	In tertiary education in the quarter given they were not in tertiary education in the previous qtr.

Model	Type	Model ID	Description
Tertiary Education	GLM	edu_txt	In tertiary education in the quarter given they were in tertiary education in the previous qtr.
Tertiary enrolled	GLM	edu_tenr	Probability of higher level of tertiary education enrolled than ever previously (age <=30)
Tertiary enrolled	Probability table	edu_tenrl	If higher level of tertiary education enrolled than ever previously, what level is it (age <=30)
Tertiary attained	GLM	edu_tatt	Probability of new highest level of tertiary education attained (age <=30)
Tertiary attained	Probability table	edu_tcoml	If new highest level of tertiary education attained, then what level is it (age<=30)

12.3 Justice Sector

12.3.1 Corrections

The proportion of time in prison, non-prison theft sentences and other sentences are stored for the previous 40 quarters, making 120 variables in total. This is sufficient for calculating the four variables used for corrections. For each successive quarter, we delete the oldest of the 40 quarters and project the newest one:

- If there was no sentence served in the previous quarter, a binomial GLM is used to project the probability that a new sentence is served in the quarter. The GLM uses a number of demographic characteristics of the individual.
  - If no, then the sentence served variables for the new quarter are set to zero.
  - If yes, then a table is used to allocate which type of sentence is served (prison, theft or other). A second lookup table is then used to allocate the proportion of the quarter served for each non-zero variable.
- If there was a sentence served in the previous quarter, a binomial GLM is used to project the probability that either the sentence continues in the new quarter or a new type of sentence starts.
  - If no, then the sentence served variables for the new quarter are set to zero.
  - If yes, then an additional binomial GLM is used to model the probability that the type of sentence being served changes. Lookup tables for the type and proportion are then used to project the new non-zero variables for that quarter.

This allows the 120 variables encoding the ten-year sentence history to be updated for the new quarter. The four variables used in the models are then re-calculated before transition and payment models are applied.

12.3.2 Offences leading to Police proceedings

Overlay models for offences leading to Police proceedings were introduced in the 2019 modelling. The decision to use an overlay (i.e. variables are model outputs only and not used in predicting other transitions or outcomes) was made to simplify the model, as corrections history already holds a lot of predictive power for transitions and was introduced into the model long before police proceedings.

Police proceeding data are used as a proxy for offences that lead to Police proceedings. This is consistent with the Justice Sector Investment Approach model. Corrections charges form a subset of police proceedings.

The modelling process is as follows:

- First, a logistic GLM is run to determine if at least one offence has been committed in the quarter. The predictors vary significantly for those who have committed an offence previously versus not as well as for those in corrections.
- If the result is that at an offence has been committed, a Poisson GLM is used to determine the number of offences committed.
- Finally, for people who commit at least one offence, the categorisation of those offences is determined using a multinomial model for the seven offence categories. This is done independently for each offence. The multinomial model with seven classes is computationally heavy so the simplifying assumption that prior offence types do not influence future offence types was made.

There are seven offence labels which consist of four types of offences, with three of them split by whether they are high or low seriousness. These are:

- **Proactively detected – high/low seriousness** – offences that the police look for in the community. Typically, these offences do not have victims e.g. drug offences.
- **Acquisitive – high/low seriousness** – offences where the offender derives material gain from the crime e.g. theft.
- **Interpersonal – high/low seriousness** – crimes that are perpetrated against an individual (excluding sexual) e.g. assault
- **Personal Sexual** – crimes that are perpetuated against an individual and are sexual in nature e.g. sexual assault.

This offence categorisation is consistent with that used by the Justice Sector Investment Approach model.

For those who do not commit an offence in the quarter but have in the past, the duration since an offence was last committed is tracked. Similarly, for those who do commit an offence in the quarter, the number of consecutive quarters in which an offence has been committed is counted.

12.3.3 Detailed list of all models used

Table 12.3 – List of justice sector models used in projection

Model	Type	Model ID	Description
Corrections event	GLM	jus_cnew	Probability corrections event in qtr given no event in previous qtr
Corrections event	GLM	jus_crem	Probability corrections event in qtr given event in previous qtr
Corrections type	GLM	jus_ccat1	Given event in qtr and previous qtr, is the combination of prison/theft/other different than previous qtr.
Corrections type	Probability Table	jus_ccat2	For those new in qtr determine which combination of prison/theft/other
Corrections type	Probability Table	jus_ccat3	Given change in combination, what is the new combination of prison/theft/other.

Model	Type	Model ID	Description
Offending	GLM	jus_offp	Likelihood of committing an offence that leads to a Police proceeding in quarter
Offending	Poisson	jus_offn	Number of offences leading to Police proceeding in quarter given they committed offence in quarter
Offending types	Multinomial	jus_offc	Categorisation of offence leading to a Police proceeding for each offence committed

12.4 Health

Health is of particular interest to MSD given Health and Disability benefits, as well as the general incidence of mental health related problems for those receiving welfare benefits.

Mortality can also be considered a health model. More information on our mortality modelling is included in Section 8.4.

12.4.1 Mental health

The definition we use in the model for a mental health event is described in Section 5.6.11.

We model Mental Health as follows:

- First, we use a logistic GLM to model the probability of an individual having any mental health event in the quarter. This model has very different parameters depending on whether someone had a mental health event in the previous quarter or not. A large proportion of people with a mental health event in one quarter have one in the next.
- For those who are determined to have a mental health event and had one in the previous quarter, we use a logistic GLM to model the probability of the individual having the same highest level of mental health event in the quarter.
- For those have are determined to have a mental health event, and either:
  - Did not have a mental health event in the previous quarter, or
  - Did have a mental health event in the previous quarter but have a different highest level of mental health event as determined by the previous model,

  we use a multinomial to determine their highest level of mental health.
- For those whose highest level of mental health event is higher than pharmaceuticals (i.e. community or inpatient) we use a logistic GLM to determine whether they also had pharmaceuticals in the quarter.
- Similarly, for those whose highest level of mental health is an inpatient mental health event then we use a logistic GLM to determine whether they also had a community mental health event.

12.4.2 Acute hospitalisations

Acute hospitalisations are defined as an unplanned admission on the day of presentation at the admitting healthcare facility[7]. Admission may have been from the Emergency or Outpatient Departments of the

[7] Ministry of Health National Minimum Dataset data dictionary

healthcare facility or a transfer from another facility. The codes defining these unplanned admissions have been given to us by the Ministry of Health.

Our modelling approach for acute hospitalisations is as follows:

- Using a logistic GLM we model the probability of an individual having an acute hospitalisation in the quarter.
- If an individual had an acute hospitalisation in the quarter, then we use a Poisson GLM to estimate the number of days spent in hospital.

It should be noted that the number of days in hospital for an acute hospitalisation may be longer than the length of a quarter, given that it is based on if the individual has had a discharge date from the hospital in that quarter.

12.4.3 Detailed list of all models used

Table 12.4 – List of health models used in projection

Model	Type	Model ID	Description
MH event	GLM	hth_mh_new	Probability of having a mental health event in quarter
MH same highest	GLM	hth_mh_srsp	Probability of having the same highest level MH event as in the previous quarter
MH new highest	Multinomial	hth_mh_srsm	If a different highest level or did not have MH in previous quarter then what was the highest level.
Pharmaceutical	GLM	hth_mh_pharm	Probability of receiving pharmaceuticals if MH event and most serious event being more serious than pharmaceuticals
Community	GLM	hth_mh_comm	Probability of community MH event if MH event and most serious event being more serious than community
Acute hospitalisation days	Poisson	hth_ah2	Number of acute hospital bed days given an acute hospitalisation event
Mortality	GLM	hth_dea	Probability of dying for those aged under 65
Mortality	GLM	hth_dea65	Probability of dying for those aged 65 and over

Note: Mortality has been included in this table for completeness. For more information see Section 8.4.

12.5 Location

12.5.1 Methodology

Our approach in earlier work was to only model an individual changing location where they were receiving a benefit or in public housing. Recognising the importance of location in other models (for example, income and health) we widened our approach to include all people in the population in the 2019 model and have applied the same modelling this year.

- For individuals not in public housing, location is updated as follows:
  - A logistic GLM is run to determine whether the region changes in the quarter
  - If the region changes, then the region is sampled from a table of probabilities. The new TLA is then sampled from a second table of probabilities.
  - If the region does not change a second logistic GLM is run to determine if the TLA changes.
  - If the TLA changes, then the new TLA is sampled from another table of probabilities.
- For clients in public housing, their region and TLA may only change if the client is projected to apply to the transfer register for rehousing. In this case, a binomial GLM gives the probability that the client applies to the transfer register. The register characteristics (including TLA) are sampled from typical characteristics of clients entering the register. If the register application is successful in the simulation, the client's TLA and region are updated accordingly.

12.5.2 Detailed list of all models used

Table 12.5 – List of location related models used in projection

Model	Type	Model ID	Description
Region	Logistic GLM	loc_dist	Probability that someone moves region (dist_grp variable)
Region	Probability Table	loc_dist_new	Which region to move to given move
TLA	Probability Table	loc_tla_new1	Allocation of new Territorial Local Authority when region has changed
TLA	Logistic GLM	loc_tla	Probability the Territorial Local Authority changed given person has not changed regions
TLA	Probability Table	loc_tla_new2	Allocation of new Territorial Local Authority when region does not change
PH preferred location	Probability Table	semi_reg_trans	Preferred location for those entering the public housing register who are currently in public housing
PH preferred location	Probability Table	semi_reg_new	Preferred location for those entering the public housing register who are not currently in public housing

12.6 Industry

12.6.1 Methodology

Industry is modelled only for those who earn income in the quarter (i.e. they are not in income band 1 – see Section 11.1.2) who are aged 65 or under.

Within the projection we simulate a value for industry for an individual after we simulate whether or not they earned income in the quarter.

We model the individual's industry, given they earned some income, as follows:

- For individuals who earned income in the previous quarter, we model whether they remain in the same industry as the previous quarter.
- For individuals who did not earn income in the previous quarter, but did in a prior quarter, we model whether they return to the same industry they were in when they previously earned income.
- For individuals who are in a different industry to where they were previously, or for those who have never previously earned income, we model which industry they are in for the current quarter.

The modelled industry is then used as a predictor for determining which income band the individual moves to and their earned income amount.

12.6.2 Detailed list of all models used

Table 12.6 – List of industry models used in projection

Model	Type	Model ID	Description
Change in industry	Logistic GLM	ind_tra	Probability that someone earning income changed industry from the previous quarter
Return to industry	Logistic GLM	ind_ret	Probability that someone who did not earn income in the previous quarter returns to the same industry they were in when they last earned income
New industry	Multinomial	ind_mul	Given someone has changed industry (including entering the workforce for the first time), which industry do they move to

12.7 Changes since the previous report

None.

E

Quality assurance and limitations

13 Quality assurance

This section describes the checking, validation and peer review processes carried to ensure the robustness of the model and results it produces. This includes input data checking and validation, model checking and validation, and projection checking and debugging. The people who have produced the model are governed by the actuarial profession's professional standards, which place strict requirements around quality of work.

Modelling complicated real-world systems over the long-term is inherently challenging. The model we have produced is very sophisticated and complex. There is undoubtedly scope for implicit and/or explicit errors to occur in the work that, without sufficient quality assurance processes, could go unnoticed. This can occur at any stage of the modelling process from sourcing data through to the interpretation of output.

We employ a range of quality assurance processes to protect the integrity of our work and ensure the results we produce are appropriate. Our team prides itself on the quality of our technical work. Quality standards covering our modelling are embedded into our everyday practice:

- Quality assurance roles are built into teamwork methods – we use formal and informal internal peer review, enjoy a collegiate culture where ideas are shared, discussed and validated, and undertake checks along the way.
- Quality assurance roles are built into team structures – we use appropriately qualified, highly skilled and experienced people, and assemble our teams carefully. Every Taylor Fry report that provides actuarial advice is authored by at least two qualified actuaries or authored by one qualified actuary and internally peer reviewed by another.
- We have customised procedures and modelling tools that allow us to quickly fit good models and assess their goodness of fit through a mix of formal statistical tests and chart-based checks. We review the predictive accuracy of models from the previous report against actual experiences and use this performance to improve models during the refitting process.
- When sub-models are chained together in projections it can sometimes lead to instability. We carry out a series of checks to help ensure robustness and quality in the result.

As well as being good work practice, adherence to sound quality assurance processes is required to comply with our actuarial professional standards – see Section 13.4.

For this project, specific attention has been applied to:

- Input data checking and validation
- Model checking and validation
- Projection checks and debugging.

13.1 Input data checking and validation

Data is the foundation of any modelling project. While adjustments can be made to models to accommodate for data deficiencies, no amount of sophisticated modelling can compensate for fundamentally flawed datasets. Therefore, ensuring that the accuracy and consistency of the data being used is appropriate for the purpose of the modelling is of paramount importance. It is also very important that the impact of any residual data deficiencies on the validity of results be understood and communicated.

Input data checking and validation is covered in Section 5.3.

13.2 Model checking and validation

The projection is built on lots of statistical models, mostly GLMs. Therefore, it is important that each model is carefully checked and validated. Model statistics relating to model fit are available on request.

There are three main checks for each model:

- Model validation
- Correct calculation of fitted values
- Reasonableness of model projections

Each of these is covered below.

In order to test that the developed models are not overfit to the data, we have also held back 10 per cent of the data in the form of a holdout data set. That is, model fitting has only occurred on 90 per cent of the data, with the other 10 per cent available for validation purposes during development.

13.2.1 Individual model validation

Our approach for the initial fitting and validating models differs depending on whether the model was previously fit or not.

13.2.1.1 Refitting an existing model

Where we have an existing model the first step is to score the old model against the new data. Any significant differences at this stage are investigated. There were some cases where this identified different data definitions from previously.

After the old model has been scored, the model is refit using the same formula as previously. Following this, any insignificant terms are removed to produce a refined version of the previous model.

Next, we add new terms. We follow a two-stage process to adding new terms. Initially, we restrict ourselves to variables that were included in the data in the previous model. Various actual vs expected charts and statistics are then used to identify where additional terms are required, and these are added to the model where significant. Statistics and charts are compared to the previous model to ensure the fit has improved. After this initial refit using existing terms, the most significant models are run in the overall simulation model to investigate any discrepancies that result when chaining models together.

The second stage involves adding new variables that were developed for the current modelling exercise. This step was not carried out for the 2023 model as no new variables were developed. Actual vs expected charts are first analysed for any of these new terms to identify which terms needed to be added for each model. These are added as main effects. Following this, interactions of these new terms are then analysed and added where necessary. Terms that are now insignificant were also dropped out of the model.

13.2.1.2 Fitting a new model

We have a routine model building process which aims to detect the main relationships in the data. An initial machine learning pass through the data is used to identify key variables. Next, we use a version of backwards selection where all key variables identified from the machine learning pass are included in the model. Following this, the model is simplified (insignificant terms are removed, and splines are used where appropriate). This process is not automatic but is carried out by a skilled modeller to ensure that modelled effects are reasonable for each variable, particularly for variable values where data is sparse. Finally, we check for interactions between key variables.

At each stage we compare goodness of fit graphs (actual vs expected charts) and statistics (gains tests) to validate the model change proposed at that point. We compare these charts and statistics with the model fit from the initial machine learning pass. Where the GLM is significantly worse, additional terms are added.

An interaction checker is run to help identify any missing interactions. In cases where the GLM is still significantly worse than the initial machine learning model a second machine learning model is used on the errors of the GLM. Information from the variable importance from the second machine learning model is used to identify where further interaction terms are required.

13.2.1.3 Reviewing process

Most of the new models were fit by an experienced modeller/qualified actuary. These were then reviewed by another experienced modeller.

We apply a number of checks as a matter of course as part of the review of any model. These include:

- Checking against a standard set of charts by quarter, duration supported by benefit/in housing and age
- Checking goodness of fit for key parameters (residual plots, actual vs expected plots).
- Statistical tests of goodness of fit (gains test)
- Checking distributional assumptions and homogeneity of residuals via residual plots (scatter plots, box plots, P-P plots)

See Appendix C.3 for more details.

13.2.2 Calculation of model fitted values

The projection relies on the calculation (or scoring) of fitted values from many GLMs. For efficiency purposes, rather than use inbuilt R projection functions, we have developed our own formula-based functionality which improves efficiency by leveraging the fact that we develop all statistical models to include continuous variables only.

For each model that is fitted and saved we include with the formula a 'model_small' dataset. This is a sample of 1,000 rows from the model fitting process in R and includes data from all predictors as well as predicted values. On gathering all the models together at the start of the projection the predicted values from our scoring functions are compared against those from the model small dataset. Any discrepancies are investigated and corrected. Note that these discrepancies are usually the result of a missing variable or unexpected missing levels in a variable, rather than any errors with the scoring functionality.

13.2.3 Reasonableness of model projections

We review model projections at both micro and macro levels. For key models, the micro level considers forecasts separately for each model and examines if the projected values for the next few years are in line with recent actual values. This is a judgemental, visual process where we are looking for instances of unexplainable change moving through time from actual values to fitted values. In general, we are satisfied if we can rationalise the fitted values based on historical values and any other known influencing factors. This process is also helpful in determining if any adjustments should be made to forecasts based on future economic expectations.

At a macro level, we examine the estimates from the projection for reasonableness. We do this in two main ways – we compare the estimates of previous models against emerging experience for an out-of-time test and secondly, we compare estimates from the current model against recent experience and assess for reasonableness- we look for discontinuities in experience or trends that are at odds with recent experience. This includes quarterly time series for:

Counts of people

- by benefit category, housing status, and register priority
- where relevant, by incapacity group, partner status and child count
- by income band, including pensioners with/without income, and by income band and benefit category
- by CNP and YJ status

- by age band and by ethnicity
- by their highest tertiary qualification attained
- deaths by benefit category, housing status, age band and ethnicity
- enrolling in tertiary enrolments, by benefit and/or housing group, age band, enrolment level
- in corrections and/or prison, at aggregate level and by benefit category
- with acute hospitalisations, at aggregate level and by benefit and/or housing group
- with criminal offences, at an aggregate level and by previous offence history, offence type and benefit and/or housing group
- by industry code
- by mental health highest event in the quarter
- by mental health highest event in the quarter and the previous quarter
- with each mental health flag (pharmaceutical, community, inpatient)
- by district group

Total payments made for

- beneficiaries by benefit category and payment type
- IRRS
- AS by housing status
- NZ Super by age band
- WFF by benefit category
- Emergency housing by benefit and/or housing group

Average

- Hospitalisation days for mental health and/or acute hospitalisations
- Number of offences for those who committed offences
- Emergency housing payments for those who received them
- Income by income band
- Income by industry code
- Income before and after transitioning to/from NOB and SUP, or aging into PEN

Probabilities for

- transitioning to SUP by current benefit category
- transitioning to NOB by current benefit category
- transitioning to Main Benefit from SUP and NOB
- entering the housing register while on/or not supported by benefit, or aged over 65
- exiting housing
- being in corrections depending on corrections history
- transitioning from and into income bands
- transition between each mental health highest interaction level
- Offences leading to Police proceedings depending on offence and corrections history

- enrolling in tertiary qualifications depending on enrolment history
- attaining tertiary qualification by attainment level

Ratios of

- income before and after transitioning to/from NOB and SUP, or ageing into PEN

Proportional distributions of

- income band and/or group by age band or starting benefit segment

Income pathways

- by benefit category,
- for NOB by age band or income band.

13.2.3.1 Reasonableness checks of the current model estimates against recent experience

As well as the comparisons of recent and projected experience by quarter discussed at the start of this section, an additional useful check is to compare the lifetime welfare and housing costs forecast by this model against those at the previous modelling date. Changes in result should be explainable– e.g. caused by data or modelling changes.

13.3 Projection checking and debugging

The MSD projection is a complex coding task, with many models linking together to produce results. It is both an extensive data and modelling task and a piece of software. Sections 13.1 and 13.2 have discussed checks and validation around data and models; here we discuss the more software-oriented tests that we carry out to ensure the projection software is working as intended.

13.3.1 Data

The projection takes a number of data sets as inputs. These include the starting cohort, new entrants and sampling populations of representative characteristics for register applicants, and new entrants. It is important that all data values be in the correct format and furthermore, that there be no unexpected missing values, and that the values that are there are as expected.

We have set up some testing functionality to check the input data to ensure that it is in the correct form. This serves as a gatekeeper to the projection – data may not be used until it passes these tests. It is important to note that the purpose of these tests does not include checking that the data is valid. This happens during the data preparation process (refer to Section 5). Rather the purpose of these tests is to check that there are no invalid data entries that would cause the projection code to fail or run incorrectly. For example, the region field must always be populated, and values must always be one of the eleven Work and Income regions from Figure 6.13. Similar tests are carried out for other variables.

13.3.2 Model checks

As noted above, we use custom-built functions for scoring all statistical models which are checked at the time of construction to ensure they work as intended. For probability tables, we check that all probabilities sum to one and ensure any large changes from the previous modelling round are explainable.

13.3.3 Unit tests of critical functions

We have set up detailed unit testing of two areas of key functionality that are used repeatedly throughout the projection code. These are:

- Scoring a probability model and (optionally) simulating a value from this model

- Sampling a value from a probability table.

There are over 100 tests in total carried out to ensure these functions are behaving as expected.

13.3.4 Development spot checks

During development we frequently spot-check the projection process to ensure that it is behaving as expected. Our code is developed with a verbose mode which can be run during the development and debugging process to look at the inputs and outputs of various projection steps in more detail.

13.4 Compliance with actuarial and accounting standards

There are currently no accounting or actuarial professional standards strictly applicable to the projection of social outcomes and associated cash flows. However, in general we carried out the estimation in accordance with standards applicable to the valuation of accident compensation liabilities.

As such, we have generally complied with the New Zealand Society of Actuaries Professional Standard No. 30 entitled “Valuations of general insurance claims”. We have also, where appropriate, complied with International Financial Reporting Standards (IFRS). Specifically, estimates of future payments incorporate an allowance for future inflation and investment return on a basis specified by the standards. However, we have not estimated nor incorporated a prudential margin as is sometimes required by such standards. In our opinion this seems unwarranted given the use to which the projection of cash flows will be put.

It is worth noting that in October 2013 the International Actuarial Association published an International Standard of Actuarial Practice 2 (ISAP 2) “Financial Analysis of Social Security Programs”. We do not believe that the standard’s intention is to cover the type of social benefit system in New Zealand; the focus appears to be on schemes with narrower scopes and elements of funding. In any event, we consider that this modelling complies with those sections of ISAP 2 that may be considered relevant.

14 Reliances and limitations

In this section we describe the risks inherent in the modelling process and the limitations these impose on how the results can be interpreted. This includes reference to the limitations inherent in using the IDI IT infrastructure.

14.1 Introduction

In preparing this report we have relied on data and other information provided in the IDI without audit or independent verification. We have carried out checks of reasonableness and consistency – see Section 5.3.2. Any known material discrepancies in the data should be reported to us so that we can consider whether this report should be amended accordingly.

There is an inherent limitation on the accuracy of estimates in this report caused by the fundamental uncertainty of attempting to predict the future. In our opinion, we have used techniques and assumptions that are appropriate, and the conclusions in reporting of results are reasonable, based on available information. However, it should be recognised that the outcomes people experience can be expected to differ from our estimates.

It is also worth noting that the outcomes we estimate are inherently complex in nature. As a simplification of reality, a model will always have limitations.

The estimation of outcomes and associated cash flows for both the current population and future entrants is subject to influences whose effects cannot be determined with accuracy. Consequently, it is a virtual certainty that the ultimate experience will depart from any estimate, but the extent of this departure is subject to uncertainty. If potential outcomes and their relative likelihood were expressed as a probability distribution, we would consider our estimates to be the mean of that distribution. In particular, the estimates provided in this report contain no deliberate bias towards over- or under-estimation.

14.2 Nature and implications of risks

14.2.1 Nature of risks

The sources of uncertainty in our model estimates can be grouped into the following categories:

- Independent (non-systemic) risk: Risks due to random variability in the incidence of outcomes and associated cash flows, despite appropriate model structure. We judge this to be a relatively small component of the overall risk.
- Systemic risk: This includes risks that, potentially, are common across more than one outcome type.
  - Risks which are internal to the modelling process, which may also be referred to as model specification risk. This risk derives from the uncertainty over to what extent the models and modelling process as a whole deviate from a perfect representation of the complex, real-life systems being modelled.
  - Risks external to the modelling process which include future changes in the environment. This uncertainty reflects the fact that, even if our modelling was perfectly correct, future legislative, policy, behavioural, demographic or economic changes may result in actual experience differing from our estimates.

It would be possible to give precise quantification of the independent risk, by combining the standard errors arising from the various sub-models built. However, given the probable size of systematic risk factors, such an estimate would likely prove misleading. Systematic risks are very difficult to estimate; however, they are mitigated by ensuring consistency in how they are treated across modelling years.

14.2.2 Potential implications of internal model specification risk for the main estimates

Model specification risk may be minimised by following good modelling practices which include robust model structures reflecting key drivers, and thorough testing of the models. However, even after following these steps, the resulting models will still be an imperfect reflection of reality. There is a real risk that future results may deviate materially from estimates due to factors excluded from the models.

By its nature, model specification risk is difficult, if not impossible, to quantify.

14.2.3 Potential implications of external risks for the main estimate

Understanding the impact of changes external to the modelling process is a key reason for conducting the modelling. Thus, external risks to the accuracy of the main estimates include:

- Future policy and operational changes
- Differences from forecasts in economic assumptions (unemployment, inflation and discount rates, rental growth etc).

We make no attempt to forecast, for example, future policy changes. We have used standard Treasury forecasts as the basis for our economic assumptions.

Understanding the sensitivity of the projection model to changes in drivers can be useful in managing the benefit system and public housing systems.

14.3 Public housing specific limitations

The public housing system is the most complex aspect of the model for two main reasons:

- Limited public housing supply means an individual's or household's future use of public housing is not independent of other individuals or households.
- Future public housing use of one household member is intrinsically linked to that of other household members

While sophisticated, our approach does simplify some important household dynamics that might materially affect the results.

14.3.1 Pseudo households for future entries

Current households are treated properly as a unit – the housing state transitions of one householder will affect the transitions of another. However future households are not linked together. This could potentially bias the projection results. For example, if the average probabilities of leaving households for non-signatory future householders are, on average, lower than they would be if we were able conditionally estimate the exit probabilities based on the linked signatory, then exit rates for non-signatories would be biased downwards. We have no grounds to believe our process is necessarily biased upwards or downwards, so consider it a reasonable approach. Nonetheless the possibility of bias is noted.

14.3.2 Household evolution

We have simplified the evolution of households within public housing. We have not explicitly allowed for the ageing of children (for instance, the number of bedrooms may change as children age from infancy, to teenage years, to adult children remaining in the house), nor the changing size and composition of notional households. We have not allowed for direct entry of adults into public housing; we always assume they enter via the register.

14.4 Inability of the projection to reflect real-world complexity

All models are simplifications of complex systems. This simplification assumes that factors not modelled remain generally stable over time. In reality, there are many factors outside the scope of the model that are likely to evolve with time. We give a few examples to illustrate the flavour of such factors below, but there are many others.

- We do not model factors such as living circumstances (aside from public housing) or access to public transport, although both have been shown to be relevant for employment outcomes. Should the mix of these factors among population of interest change substantially, we would expect experience to differ from estimates.
- Society's attitude to different government services might evolve over time.
- Natural disasters such as the Christchurch earthquakes and the COVID-19 pandemic have significant effects on government service use. The timing and severity of external events of this type are not able to be predicted with any certainty. We do not attempt to allow for such future events in our estimates.

Such issues require us to consult closely with MSD to ensure we understand recent factors that affect the models as they become apparent. Not modelling these factors does not imply a failure of the projection. It still provides important feedback and can allow for significant events and trends as they occur.

14.4.1 Impact of COVID-19

We note that the COVID-19 pandemic continues to add uncertainty to our estimates, albeit to a lesser extent than the 2021 model. While the effects on the outcomes we estimate are expected to be short-term, the extent of the effects and exactly how long they will persist are unknown. The effects also hard to distinguish from broader labour market effects that may or may not have been present in the absence of the pandemic.

14.5 Other specific limitations of the projection

There are significant implementation challenges associated with the following issues:

- The use of simulation to generate projections: We estimate the 'noise' typically associated with simulation projections is negligible at an aggregate level, but it is potentially significant at the cohort and individual level. Extra simulations may be required for small subgroups of interest.
- Changes to the systems: Changes, such as the changes to benefit types in 2013 can cause practical challenges e.g. difficulties in reconciliation between the old and new systems.
- Impacts of COVID-19: We know that because of COVID-19 transitions rates will at least temporarily be different from our modelled rates. We have adjusted our projected rates temporarily based on the Treasury's HYEFU unemployment rate assumptions and experience during the pandemic so far. These assumptions themselves are uncertain.
- Data matching limitations: There are inherent technical limitations to how well the datasets are matched in the IDI. We use these variables aware that a small, but material, number of false positive and false negative matching occurs.
- Public housing data: This data is of lower quality for longitudinal modelling. This inherently limits the degree to which we can accurately model future public housing pathways.

None of the items above undermine the accuracy or usefulness of the projection. We raise them primarily so that MSD is aware of some of the issues likely to arise in future.

14.6 Use of the Integrated Dataset Infrastructure

While the data available in the IDI have significantly enhanced the modelling, the available IT infrastructure has placed some limitations on the project. There are three limitations, in particular, that we note:

- SQL server resources shared among many users in the IDI. All of our initial data preparation is performed using the SQL server and data from the SQL server is called during our model fitting process. When there are many users using the SQL server the speed of any queries slows significantly.
- Storage space on the file and SQL servers are constrained. There is a hard cap on the storage space available in the SQL server and storage space on the file server is costly. This makes it expensive to store the output from multiple simulations concurrently and to store datasets and outputs from prior years. The cap on storage space in the SQL server makes it necessary to shift data to and from the file server which is time consuming and makes it difficult to keep track of datasets.
- Within the projection code itself, we have made a number of design choices to reduce memory usage. The main one of note is the bucketing of the general population into groups of a variable size to reduce total data set size within the projections (see Section 8.6). This means that the variability of the wider population within the simulations is reduced somewhat since buckets share the same characteristics. This is mitigated somewhat by the fact that most of these characteristics are sampled at the start of the projection, so different simulations will have different characteristics associated with each bucket.

14.7 Disclaimer

All results that use IDI data are subject to the following disclaimer:

These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. For more information about the IDI please visit https://www.stats.govt.nz/integrated-data/.

Access to data used in this study was provided by Stats NZ under conditions designed to give effect to the security and confidentiality provisions of the Data and Statistics Act 2022. The results presented in this study are the work of the author, not Stats NZ or individual data suppliers.

The results are based in part on tax data supplied by Inland Revenue to Stats NZ under the Tax Administration Act 1994 for statistical purposes. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

F

Appendices

Appendix A Glossary

The following table gives definitions for common acronyms and terms used in this report.

Table 14.1 Acronyms and terms used in this report

Term	Definition
ABP	Average benefit paid per quarter to clients in receipt of a benefit that quarter.
Applicant	An Applicant is the primary household member in a public housing application while on the public housing register.
AS	Accommodation Supplement (and related assistance)
Average future lifetime cost	Refers to the expected future benefit payments to a client up to age 65, including inflation and discounting. Sometimes shortened to 'average lifetime cost' or 'average cost' but excludes benefit payments to the client made before the projection date.
AWE	Average Weekly Earnings
BEN	Receiving a Main Benefit, this includes Jobseeker support, Sole Parent Support, Supported Living Payment, Young Parent Payment and Youth Payment.
Board	Community or Local Board - geographical subgroup of territorial local authorities.
CCS	Childcare subsidy (including OSCAR payments to clients)
CDA	Child disability allowance
CHP	Community Housing Provider - a housing provider (other than Kāinga Ora) that provides social rental housing and/or affordable rental housing.
CPI	Consumer price index
CNP	Care and protection for children
DA	Disability allowance (and related assistance)
EB	Emergency benefit (included in Jobseeker Support benefit)
EI	Employment Intervention
EMS	Employee Monthly Schedule
GBM	Gradient Boosted Machine
HCD	Health condition, disability (sub-set of both Jobseeker Support and Supported Living Payment beneficiaries with reduced work obligations).
HS	Hardship Benefits
HYEFU	Half-year Economic and Fiscal Update

Term	Definition
EI	Supplementary Assistance - Employment interventions (including training provided as supplementary assistance).
EMH	Emergency housing
HS	Non-recoverable hardship assistance
HUD	Ministry of Housing and Urban Development
IDI	Integrated Data Infrastructure – Research database containing microdata about people and households from a range of government agencies, surveys and non-government organisations.
IFRS	International Financial Reporting Standards
IRD	Inland Revenue Department
IRR	Income-related rent – IRR is calculated based on a client's assessable income and their household type. Public housing providers (Kāinga Ora and CHPs) then charge this rate as rent to the client (market rent = IRR + IRRS). If the calculated rate of IRR is higher than the market rent for the property, the housing provider will charge no more than the market rate as rent for the property.
IRRS	Income-related rent subsidy – a top up payment to housing providers to bridge the difference between the income-related rent a client pays and the market rent of the property. Market Rent = IRR + IRRS.
JS	Jobseeker Support – new benefit type introduced July 2013 (replaces Unemployment Benefit and Sickness Benefit, and partially replaces Domestic Purposes benefit). We sometimes refer to people receiving JS as Jobseekers, or JS.
Loans	Covers all cases where a client can become indebted to MSD, i.e. via overpayments of benefits or assistance (inadvertently or through fraud) or via recoverable assistance (including both benefit advances and other recoverable assistance).
Market Rent	The average level of rent being paid for similar properties in the same area. Market Rent = IRR + IRRS.
MCA	Multi-category Appropriation
MoE	Ministry of Education
MSD	Ministry of Social Development
NCEA	National Certificate of Educational Achievement
Neither	Not in a public housing place and not receiving Accommodation Supplement. Sometimes referred to as NIL.
Net loans cost	The future payments for the cost of loans after allowance for recoveries.
NEET	Not in Education, Employment, or Training

Term	Definition
NIL	Not in a public house and not receiving Accommodation Supplement. Sometimes referred to as 'Neither'.
NOB	Not supported by a benefit (in a given calendar quarter)
NOMB	Not supported by a main benefit (in a given calendar quarter) but still receiving some benefit system support – a supplementary benefit or OB.
NZQF	New Zealand Qualifications Framework
NZ Super	NZ Superannuation – A non means tested payment to New Zealanders aged over 65 who meet the residency requirements, also includes the Veterans Pension.
OB	Orphan and unsupported child benefits
Offense	Police Proceeding – An offense reported to the NZ Police who then took out proceedings against the individual.
OSCAR	Out of School Care and Recreation subsidy to providers.
OTH	Other benefit, referring to those clients not on a key benefit, includes supplementary assistance, but not including JS-SH (student hardship), CCS, EI and HS.
Overpayments	Payments (benefit or assistance) where a client is inadvertently paid more than their entitlement. In the projection overpayments include those due to fraud.
PH	Public Housing – clients are considered in public housing if they reside in a property managed by Kāinga Ora or a Community Housing Provider, they may be paying income- related rent or market rent.
PREFU	Pre-election Economic and Fiscal Update
Recent exit	Recent housing or register exit – a client who is currently not in a public housing place or on the register but has been in the last 12 months. Recent benefit system exit – a client who is currently not receiving a benefit but has been in the last 12 months
Recoverable assistance	In this report recoverable assistance includes benefit advances and recoverable assistance.
Recoveries	Repayments of overpayments and recoverable assistance to MSD.
REG	Register – refers to the public housing register, used to manage applications for public housing.
Region	A geographical grouping by MSD of New Zealand into 11 regions.
Signatory	A signatory in a household is a person who signs the tenancy agreement and whose income is included in the households' income calculation. Refer to 'Tenant'.
SNG	Special Needs Grant

Term	Definition
SLP	Supported Living Payment – new benefit type introduced July 2013 (replaces Invalid's Benefit and Domestic Purposes Benefit – Care of the Sick and Infirm).
SPS	Sole Parent Support – new benefit type introduced July 2013 (partially replaces Domestic Purposes benefit). We sometimes refer to people receiving SPS as Sole Parents, or SP.
SUP	Clients receiving supplementary benefits (Tier 2 or 3), but no main benefit.
SWA	Social Wellbeing Agency
SWN	Social welfare number
SAS	System allocation score (derived from a household's public housing application responses)
Tenant	Clients are sometimes referred to as tenants where they reside in a property managed by Kāinga Ora or a Community Housing Provider, they may be paying income-related rent or market rent.
TLA	Territorial Authority where the individual lives. For those in the Auckland Territorial Authority this is instead the Community Board.
Transfer	This term is used to describe a client who transitions from one benefit type (or segment) to a different benefit type (or segment).
WEP	Winter energy payment
WFF	Working for Families tax credits. In this report this covers the family tax credit, minimum family tax credit, in-work tax credit and parental tax credit (now replaced by the Best Start tax credit).
WR	Work-ready (sub-set of Jobseeker Support beneficiaries with work obligations)
YJU	Youth Justice
YP	Youth Payment
YPP	Young Parent Payment

Appendix B Methodology for projecting regional unemployment rates

B.1 Historical series

Our projection models use a seasonally adjusted unemployment rate for New Zealand and its regions. Regional rates are only available in raw form, i.e. not seasonally adjusted. Therefore, for consistency in our modelling process, it is necessary to first produce seasonally adjusted series of regional unemployment rates. We also remove some of the quarterly volatility via smoothing.

Our approach to producing adjusted regional unemployment rate series is as follows:

- Source raw data from Statistics NZ
- Calculate de-seasonalisation factors, taken as the average amount that quarter of year is above or below the average for a five-year moving window centred at that date. For example, the 1991Q2 de-seasonalisation factor is the average unemployment rate for Q2 in '89, '90, '91, '92, and '93 compared to the overall average in those five years
- Centre the de-seasonalisation factors so that each rolling year of factors is centred at 100%
- Use these centred de-seasonalisation factors to produce seasonally adjusted time series
- Smooth the time series by using neighbouring quarters:

$$UE(t) = 0.25\ UE(t-1) + 0.5\ UE(t) + 0.25\ UE(t+1)$$

B.2 Projection series

The following approach is used to derive regional forecasts:

- Find regional weights using the average total labour force over 2022/2023.
- Assume the quarters from 2015Q3 through to 2019Q2 represent a period of stable unemployment and calculate the average unemployment in each region over this period.
- Calculate the difference between the regional average and national average over that period. These differentials are used in the regional long-term rate assumption.
  - Currently Treasury uses 4.25% as the national long-term unemployment rate. For example, a differential of +2.4% was calculated for Northland (over 2015-2019), so the Northland long-term rate is 6.65%.
- Mirror the Treasury projection shape for each region, taking the unemployment rate from the current level to the long-term average rate over five years.
- Add a correction factor to each future quarter, to ensure that the weighted average unemployment rate equals that used at the national level.

Appendix C Generalised linear models

Most of the models used in the valuation are generalised linear models so we give a brief overview of the theory behind these models here, our processes for fitting GLMs and the diagnostics we use to validate the model fit.

C.1 Theory

A generalised linear model ('GLM') is a generalisation of ordinary least squares regression that is able to deal with non-normally distributed response variables. Given a response variable y and a set of independent variables or predictors $x_1, x_2, \ldots, x_n$ a GLM models the dependency as:

$$y = h^{-1}\left(\sum_{i=1}^{n} \beta_i x_i\right) + \varepsilon_i$$

and

$$E(y) = \mu = h^{-1}\left(\sum_{i=1}^{n} \beta_i x_i\right)$$

where

$h^{-1}()$ is the link function
β_i (i=1, 2, ..., n) is the parameter corresponding to the independent variable x_i
ε_i is an error term

Note that

$$\eta = \sum_{i=1}^{n} \beta_i x_i$$

Is referred to as the linear predictor and that the GLM may be written as:

$$y = h^{-1}(\eta) + \varepsilon_i$$

Thus, a GLM consists of three components:

- A probability distribution
- A link function
- A linear predictor

Probability distribution

In the equations above, the error term ε_i is determined by the probability distribution of the response variable. Common distributions that may be used include:

- Normal
- Poisson (including the over-dispersed form where the variance is given by $Var(y) = \varphi\mu$ for $\varphi > 1$)
- Gamma
- Binomial

The choice of distribution is informed by the response variable. For example, counts are naturally modelled by a Poisson distribution while strictly positive continuous quantities may be appropriately handled by a Gamma distribution. The Poisson distribution (particularly in its over-dispersed form) is also useful for modelling continuous positive quantities particularly when they may take the value zero. Probabilities may be modelled using a Binomial distribution.

Link function

The link function $h^{-1}()$ gives the relationship between the mean of the distribution and the linear predictor. There are many possibilities for the link function including (but not limited to):

$h^{-1}()$ gives the relationship between the mean of the distribution and the linear predictor. There are many possibilities for the link function including (but not limited to):

- Identity link: $h^{-1}(\eta) = \eta$
- Log link: $h^{-1}(\eta) = \exp(\eta)$
- Logit link: $h^{-1}(\eta) = \exp(\eta)/(1 + \exp(\eta))$

It is usually convenient to choose a link function which matches the domain of the link function to the range of the response variable's mean. In other words, if a response must be positive (for example, an average benefit payment), then a log link will ensure that the fitted value μ is positive. If the modelled quantity is a probability (for example, the probability of transitioning off benefit in the next quarter), then the logit link ensures that the fitted value lies between 0 and 1, as probabilities must.

Linear predictor

The linear predictor is the quantity which incorporates the information about the independent variables into the model and is typically denoted by η. η is expressed as a linear combination of unknown parameters β_i and independent variables x_i (i=1, 2, ...), which are known.

In all cases, once the probability distribution and the link function have been selected, the linear predictor needs to be constructed. The steps to doing this include some or all of:

- Identify the list of independent variables or predictors (x_i) to be considered.
- Using data exploration, modelling techniques, statistical tests and prior knowledge, identify those x_i that are useful for predicting the response variable. Note that this may include functions of the predictors, rather than the raw predictors themselves.
- Estimate the parameters β_i using GLM software.

References

The following books give a complete introduction to GLMs:

- McCullagh P. and Nelder J. (1989). Generalized linear models, second edition. Chapman and Hall, London UK.
- Dobson A. J. (2002). An introduction to generalized linear models, second edition. Chapman & Hall/CRC, Florida USA.

For a discussion on the application of GLMs in contexts similar to the modelling of the MSD benefit liabilities (e.g. claim size and claim numbers modelling in insurance), the following papers provide some starting points.

- England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. British Actuarial Journal, 8 443-544.
- Haberman, S. and Renshaw, A. E. (1996). Generalized linear models and actuarial science. The Statistician, 45 407-436.
- Mulquiney, P. and Taylor, G. (2007). Modelling Mortgage Insurance as a multi-state process. Variance 1, 81-102.
- Taylor, G. and McGuire, G (2004). Loss reserving with GLMs: a case study. Casualty Actuarial Society Discussion Paper Program 2004. Available at http://www.casact.org/pubs/dpp/dpp04/04dpp327.pdf

C.2 Process for fitting a GLM used in this work

As noted above there are three components to a GLM:

- A probability distribution
- A link function
- A linear predictor

Each of these must be selected as part of the model fitting process.

Distribution and link

The probability and link function are usually determined by the quantity being modelled. For example, many of the models used are probability models and therefore the binomial distribution with a logit link are obvious choices for the respective GLMs.

Models of payments should produce positive or zero values, so a log link ensures this (zero values being approximated by very small positive numbers). Furthermore, a log link leads to a multiplicative model which is a desirable feature. Distribution-wise, the over-dispersed Poisson is often a good choice since it can model zero payments (unlike a Gamma distribution which requires strictly positive observations).

While the type of observation may suggest a particular distributional form, it is important to validate that choice (see below). We note, however, that estimates of the mean tend to be reasonably robust to choices of distribution. Most of the projection code only uses the mean values from the micro models so there is some tolerance to deviations from the assumed distributional form. This statement that the projection code uses mean values may seem at odds with the simulation approach, but it is the case. Estimated transitional probabilities are calculated from the various GLM models and it is these mean probabilities that are used, rather than sampled probabilities from the distributions with those means and variances implied by the distributional assumptions. Likewise, when estimating benefit and housing-related payments, the mean payment values are used rather than sampled values from their distribution.

Linear predictor

The first step is to carry out what is now often referred to as feature engineering – decide what set of features to use for consideration for inclusion in the models. Machine learning algorithms such as gradient boosting machines or random forests are often useful to whittle down a long list of variables into a shorter list. Furthermore, prior and expert knowledge is also useful, particular in cases where functions of variables are needed. For example, corrections history over the last 10 years is used in the model but the actual features used are based on four variables summarising correction type and duration over the most recent year or all 10 years.

This type of exercise is not repeated from year to year, but was carried out at the start of the welfare modelling process and later with the addition of public housing to identify a common set of features to be used across all models (note that all features are not necessarily in each model). Incremental additions occur each year as new variables come on-line.

When fitting an individual model, we typically use a version of backwards selection where all variables are initially included in the model and then step-by-step, a skilled analyst simplifies the model by removing insignificant terms and by adjusting the form of the predictors where required. The significant predictors or independent variables may be used as follows:

- In their raw forms: For example, gender with two levels F and M.
- As categorical groupings of the original variable: For example, age may be banded into a number of groups (<18, 18-29, 30-39 etc). However, in this specific case, any categorical variables were replaced in the final model by indicator functions (next bullet) for efficiency gains in the simulation code.
- As indicator functions depending on the value of the original variable where one condition is assigned the value 1 and the complementary position 0: For example, letting I(age ≥ 30) be 1 for age ≥ 30 and 0 otherwise would fit a step term at age 30.

- As a spline for underlying raw predictors which are numeric or ordinal (e.g. age, benefit quarter, duration supported by benefit): The dependency of a linear predictor on duration could be modelled (if appropriate) by a combination of several line segments. For instance, if the linear predictor varied in a linear fashion with duration with one slope from duration 1 to 4, a different slope from 4 to 12 and a third slope from 12 onwards, then using three line pieces (1-4, 4-12 and 12+) would capture this dependency. The points 4 and 12 where the resulting fitted spline bends are referred to as knot points.
- As interaction terms: All of the above may be used as interaction terms. For example, a duration effect may be well fitted by one spline for those aged under 30 and another for those aged 30 and above. This could be accommodated by interacting the spline with the I(age ≥ 30) term.

In principle, backwards selection implies that all possible significant terms are included in the model at the start. However, in practice, it is usually not possible to ensure all possible interactions are included. Our general approach to interactions is to include any interactions at the start that are expected a priori. Later we use diagnostic tools to check for evidence of poor fitting which may suggest the need for other interactions. These are then added as required until a satisfactory fit is obtained.

C.3 Model diagnostics

The model diagnostic tools used are primarily visual in nature. We examine plots of residuals and of actual and fitted values. These plots yield information on the quality of the fit and/or the appropriateness of the distributional assumptions.

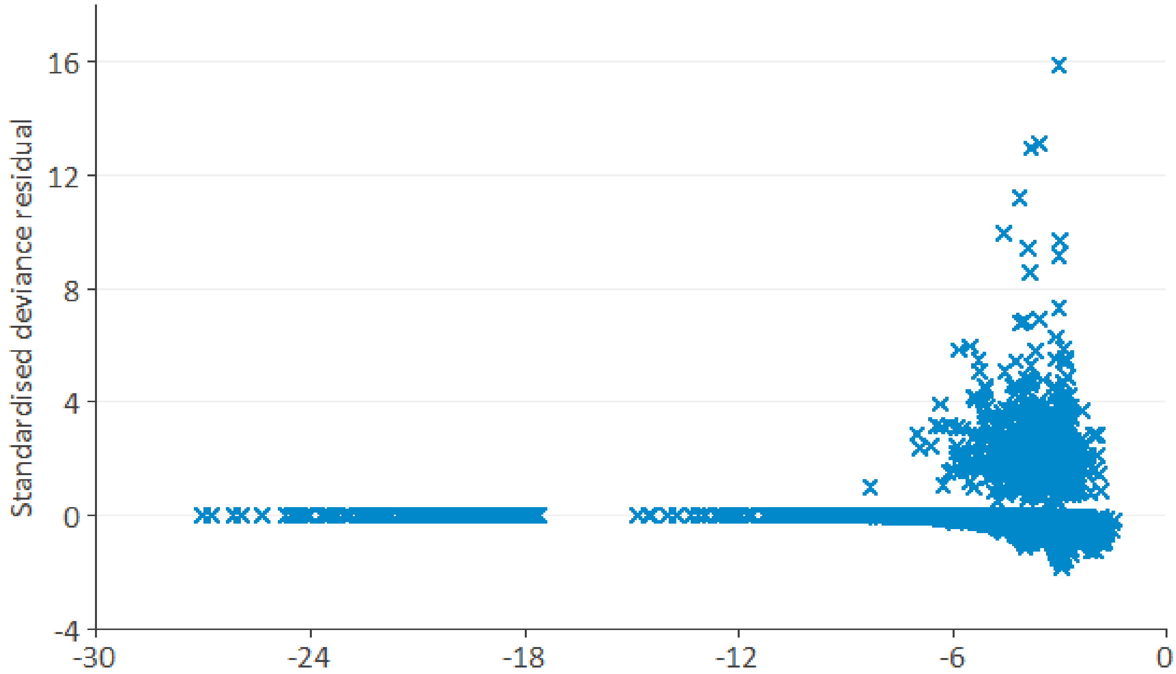
Residual scatterplots

Scatterplots of standardised residuals against linear predictors or fitted values, or against some of the independent predictors are used as a model validation tool. If a model fits well, then the residuals should be homoscedastic, i.e. display as a cloud of points centred around zero. Furthermore, standardised deviance residuals (refer to standard statistical texts such as McCullough and Nelder above for definition) should be normally distributed in a well-fitting model.

An example of a residual scatterplot is given below. Other visual summaries may be helpful to assess the distribution of the residuals such as a graph of the quartiles of the residuals, or similarly, box plots of the residuals for different levels of a variable.

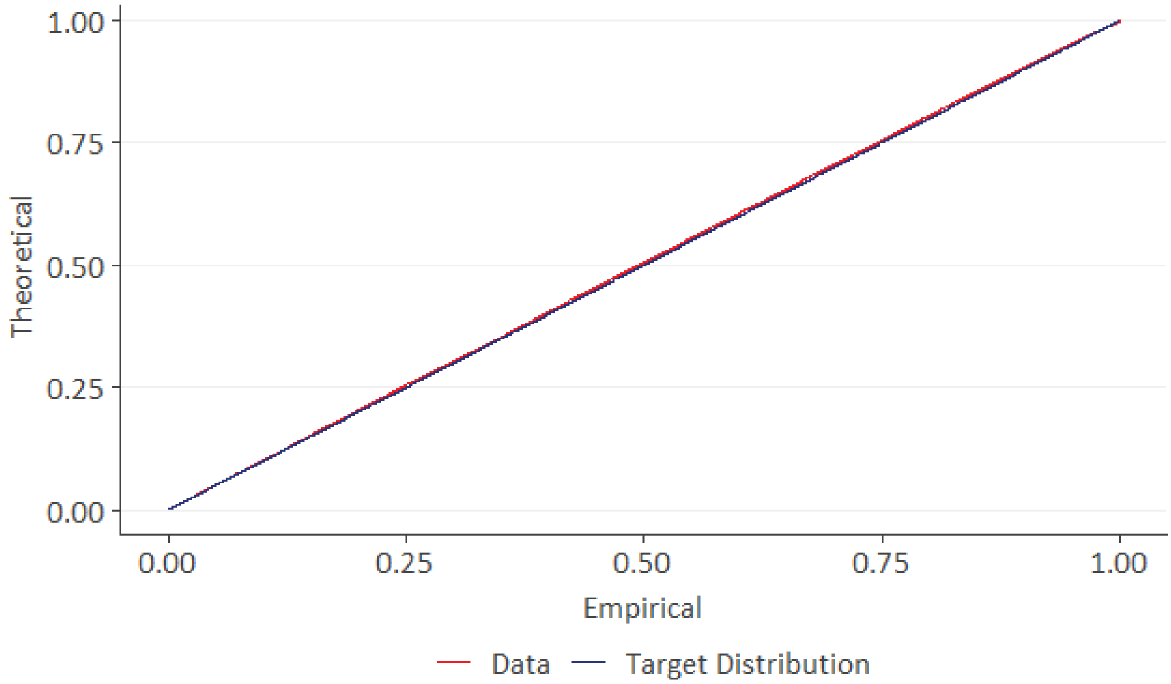
Heteroscedasticity (for example where residuals fan out) may be evidence of poor model fit or poor distributional assumptions.

Figure C.14.1 – Example of a residual scatter plot



A probability-probability (P-P) plot is also a useful way to look at the distributional assumptions. Here the theoretical cumulative probability distribution is plotted against the empirical values. If the selected distribution is appropriate, then the plotted line should be close to a 45-degree line (the blue line below). Deviations from the straight line indicate areas of poor fit. For example, if the distribution assumed has lighter tails than the data, then the P-P plot will deviate from the straight line at both extremes. The P-P plot is similar to the more well-known quantile-quantile (Q-Q) plot.

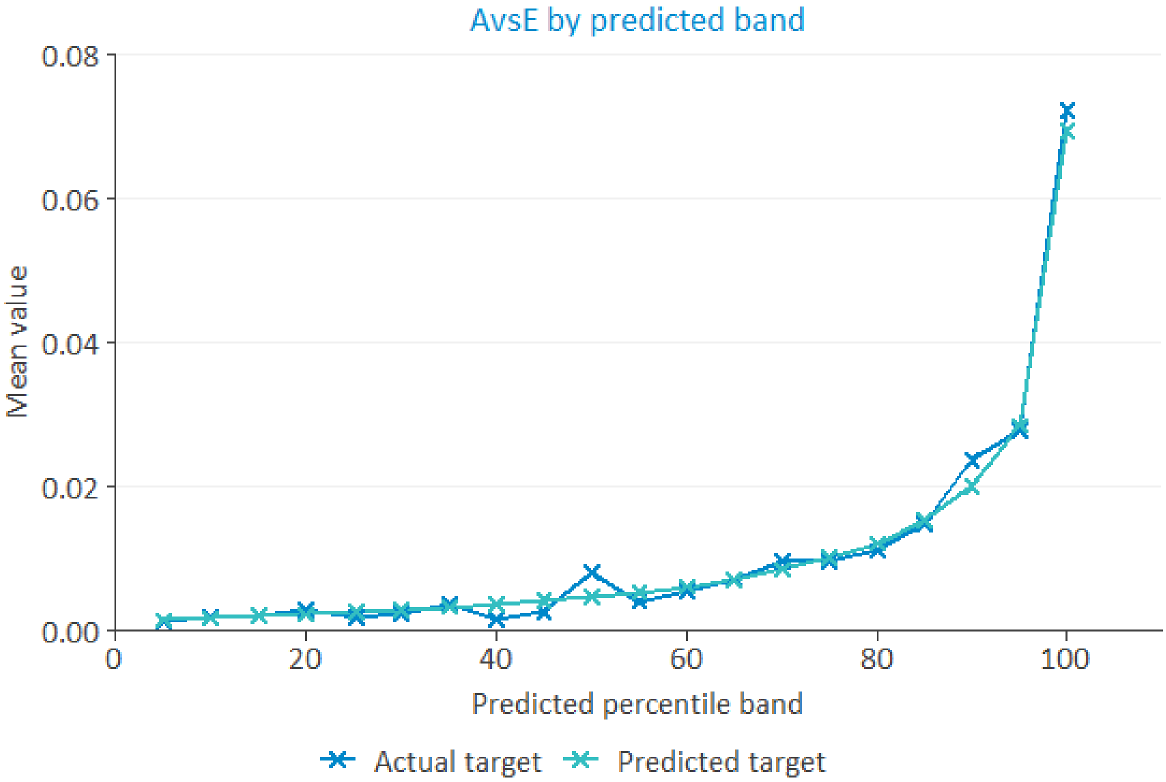
Figure C.2 – Example of a P-P plot



Actual vs expected predictor plots

A plot of the mean target and predicted values by predicted bands was used to assess high level goodness of fit. The bands are percentile groups of the model predictions, in ascending order. The chart shows how well the target and predictions correlate, for different values of the prediction. A good model will correlate well (meaning the lines are close together) and will have good discrimination (meaning the lines have a steep slope).

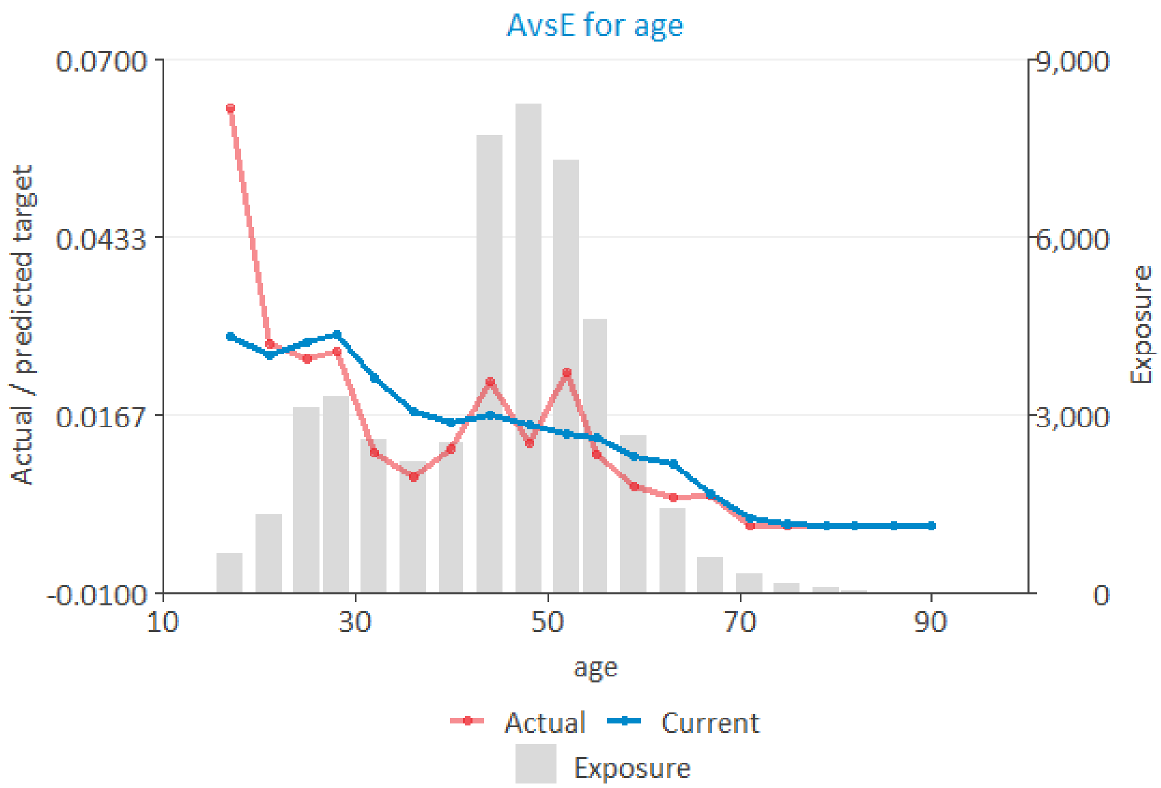
Figure C.3 – Example of an actual vs expected predictor plot



Finally plots of actual vs expected values are very useful for assessing goodness of fit. These plots display the sum of actual values and the sum of fitted values for different levels of a predictor and are a valuable tool in assessing model fit. We usually display the exposure measure in these graphs as well so that differences in actual and fitted values can be put into context. These plots can be used for the entire modelling data set, or for subgroups of interest. In predictive models, looking at the quality of fit in recent quarters is particularly useful.

An example of this type of plot are shown below which example compares the actual and predicted values by age.

Figure C.4 – Example of an actual vs expected plot



TAYLOR FRY