Micro-management: Some unsolved problems in microsimulation models

Hugh Miller

October 2022

Agenda

0. Introduction
1. Missing variables
2. Latent / hidden variables
3. Component model selection
4. Dealing with dimensionality
5. Continuous variables
6. Validation
7. Model size and speed
8. Data time windows

Introduction

Despite heavy involvement in a number of projects, there are still many questions around best practice for large actuarial microsimulations.

Much of these comments come from a position of humility and/or ignorance. There are probably good ideas addressing specific problems out there – ‘unsolved’ mat actually mean ‘incomplete literature review’

Various models built over the last decade, including:

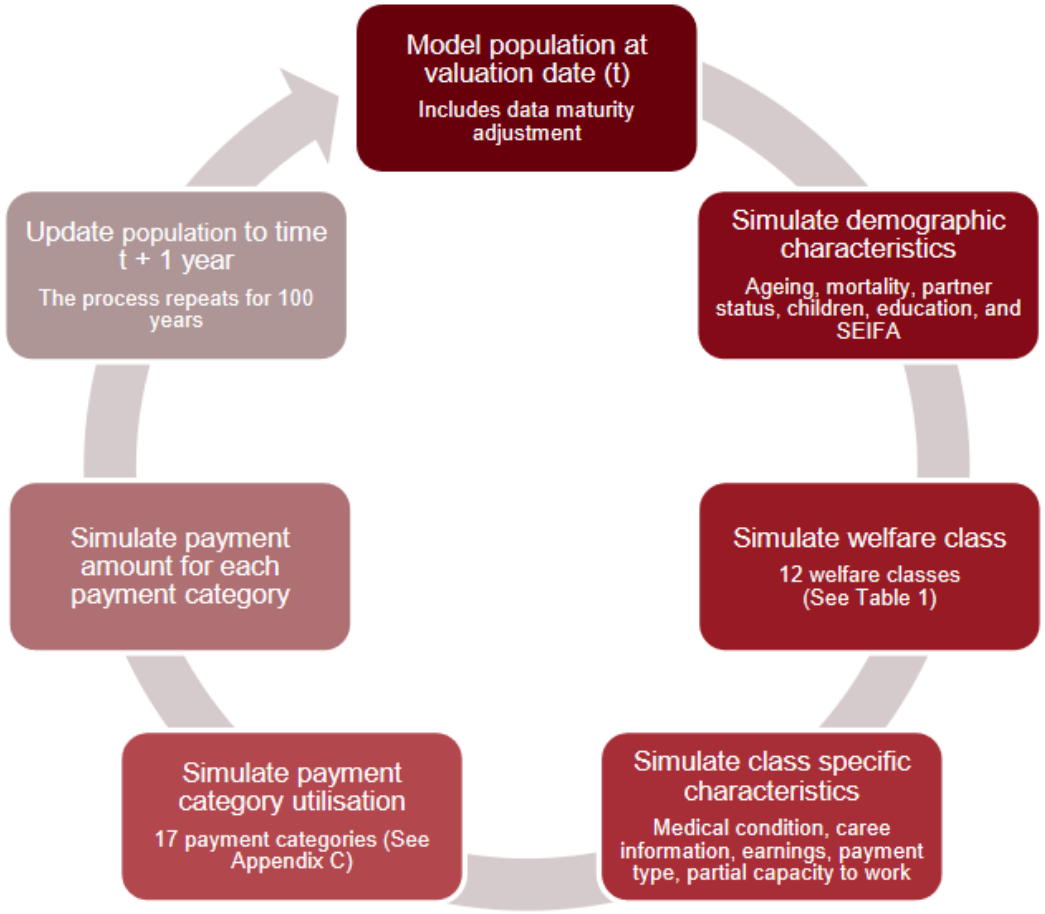
- NZ welfare and housing (now general social outcomes, from 2012)
- CMW Priority Investment Approach to welfare (2016)
- CMW Department of Veterans Affairs
- NZ Justice
- NZ vulnerable children
- NSW linked services (from 2018), with vulnerable children and families emphasis
- NSW OOHC leavers (2018)
- Others

What is a microsimulation?

For concreteness we give the example of the Priority Investment Approach to welfare model

- A person-level projection until death, annual timesteps
- Many dynamic variables. Some central (e.g. welfare state), others necessary to update as predictors (e.g. education level, SEIFA, household information)
- A range of fixed variables (e.g. gender, intergenerational information)
- Each time step is a series of component models, usually set up as a traditional learning problem (E.g. the probability of moving from welfare class 2 to class X next year, given current characteristics) with historical data to estimate (E.g. take the rates over 2017 to 2019).
- Individual person can be projected independently, perhaps multiple times.
- Metrics of interest derived (e.g. expected $ of welfare, expected years on various benefit types)

Update cycle, PIA model 2020



A selective origin story

The NZ Investment Approach

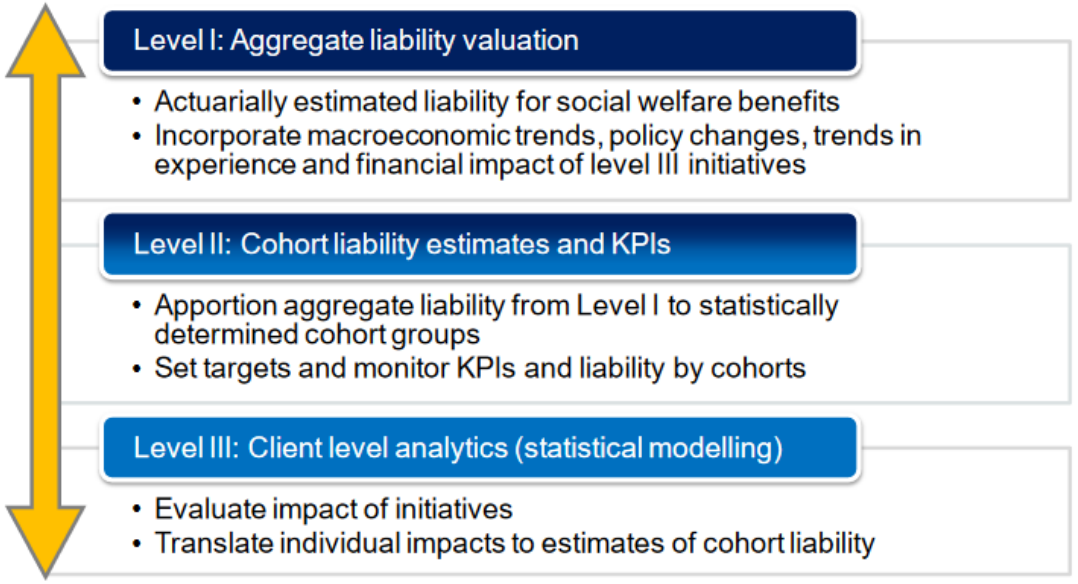
One of the first models came from the NZ Investment Approach to welfare. It required analytics and modelling cycle to track progress in reducing long-term benefit receipt through improved employment opportunities.

Original plan had three levels of modelling. Microsimulation was one potential method for doing the middle layer (cohort liabilities). The top level was envisaged for a more traditional aggregate valuation and the lowest level more standard analytics.

Level II ended up swallowing Level I and a lot of Level III. Mainly because Hugh thought it was a good idea.

Microsimulation seems to bridge the gap between aggregate long-term valuations and more agile (shorter-term) analytics

Proposed analysis framework, NZ Feasibility study



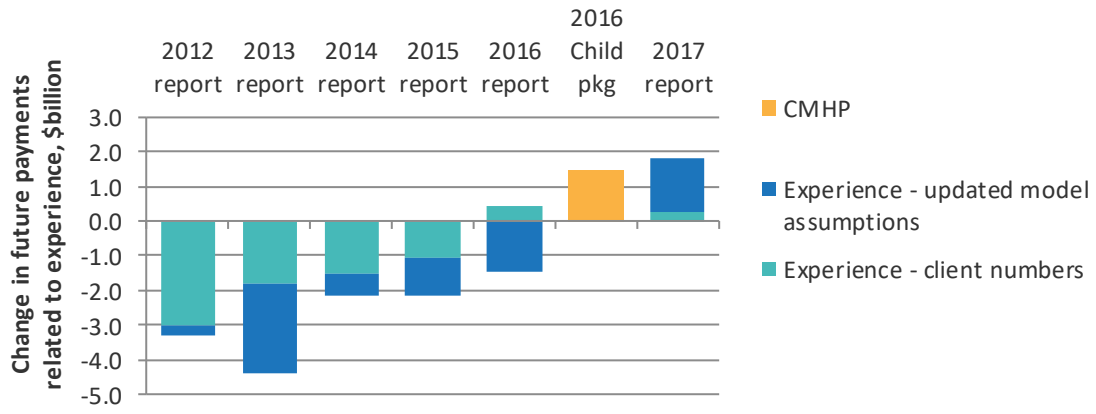
https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/evaluation/taylor-fry-ia-feasibility/taylor-fry-feasibility-of-an-ia-for-welfare-report.pdf

A selective origin story (2)

Models have proven pretty powerful:

- Longitudinal forecasts show the concentration of cost while retaining the ability to understand different risk factors and cohorts.
- Genuinely enables ‘investment’ thinking
- Annual attribution of change
- Extensibility
- Insight generation

Cumulative analysis of change



Not without criticism:

- Too focused on answering a specific question in heavy detail while ignoring others (e.g. managing a fiscal liability compared to understanding economic wellbeing).
- General concerns on validity of long-term chaining
- Time-consuming to build and maintain – specific questions can be answered in alternative ways quicker
- As a general purpose prediction tool, can be a hammer looking for a problem
- Overreliance on observational data (versus causal inference)

1

Missing variables

1. Missing Variables

The problem

Suppose we have a variable z_t (e.g. Education level) with lots of missing values. We may have two cases:

- **Case A – Missing at random, not tied to response.** Missingness is depends on other predictors (e.g. age and gender), but not on the response or the value of the education variable itself
- **Case B – Missing at random, tied to response.** Missingness is depends on other predictors (e.g. age and gender), and also our target (e.g. probability of moving benefit state)

To add complexity:

- Longitudinal (so z_t must relate to z_{t+1} for a person)
- Dynamic (can evolve, and will need to be handled by the education-specific evolution model)

Age	Gender	EDU	Current benefit	Exit welfare ?
30	F	Yr12	JS	N
24	M	Uni	JS	Y
20	M	??	JS	Y
23	F	??	JS	Y
30	M	Cert 4	JS	N

1. Missing Variables

Standard responses are unsatisfactory

- **Impute before modelling and projection.** This will dilute the signal and create bias. For example, if university-educated people exit welfare faster, and we allocate many people without degrees to the university bucket, the average exit rate will drop
- **Model before imputation, but impute before projection.** Welfare transition models might have a education missing flag built in to help create unbiased signals. But in case B, imputing after modelling will induce a bias (e.g. if missings tend to exit faster, then overall exit rate is too low if projected without missings)
- **No imputation.** Likely to induce bias if missing status is non-stationary over the lifecourse (e.g. if lots of missing for people aged 18-20, then 5 years into the projection have an unreasonable number of 23-25 year olds with missing education.

Model

Age	Gender	EDU	Current benefit	Exit welfare ?
30	F	Yr12	JS	N
24	M	Uni	JS	Y
20	M	Yr12	JS	Y
23	F	Uni	JS	Y
30	M	Cert 4	JS	N

Age	Gender	EDU	Current benefit	Exit welfare ?
30	F	Yr12	JS	N
24	M	Uni	JS	Y
20	M	??	JS	Y
23	F	??	JS	Y
30	M	Cert 4	JS	N

Age	Gender	EDU	Current benefit	Exit welfare ?
30	F	Yr12	JS	N
24	M	Uni	JS	Y
20	M	??	JS	Y
23	F	??	JS	Y
30	M	Cert 4	JS	N

Project

Age	Gender	EDU	Current benefit	Exit welfare ?
30	F	Yr12	JS	N
24	M	Uni	JS	Y
20	M	Yr12	JS	Y
23	F	Uni	JS	Y
30	M	Cert 4	JS	N

Age	Gender	EDU	Current benefit	Exit welfare ?
30	F	Yr12	JS	N
24	M	Uni	JS	Y
20	M	Yr12	JS	Y
23	F	Uni	JS	Y
30	M	Cert 4	JS	N

Age	Gender	EDU	Current benefit	Exit welfare ?
30	F	Yr12	JS	N
24	M	Uni	JS	Y
20	M	??	JS	Y
23	F	??	JS	Y
30	M	Cert 4	JS	N

1. Missing Variables

Potential solutions

Could a more sophisticated imputation stage be used that used available information, **including welfare transitions and longitudinal patterns**, to estimate education

Key questions:

- How does validation change if the 'target' is used in the imputation stage? Is that even allowed?
- In case B, where the missingness is telling us something, does anything need to be preserved?
- How does imputation interplay with monitoring setups (e.g. differences in Actual vs Expected by education level distorted by changes in missing status?)

2

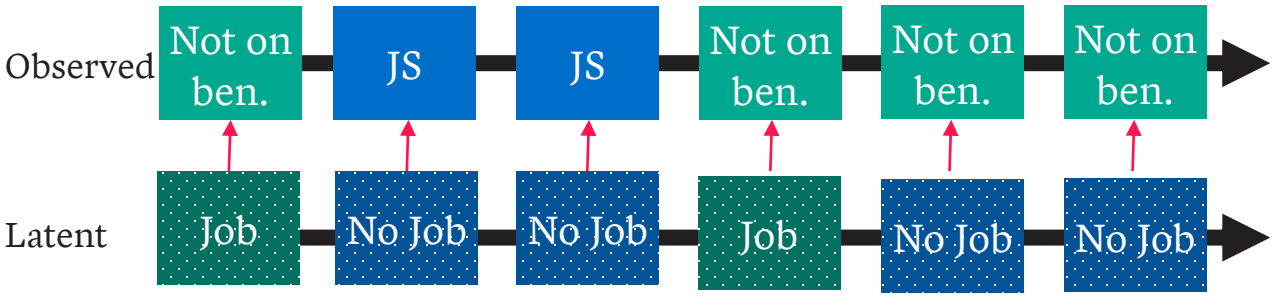
Latent variables

2. Latent variables

The problem

Often the variable we observe is not the variable we are interested in. For example, 'not on welfare' might be a proxy for 'has a job' or 'is in a financially secure household', but in reality is only a proxy for it.

This can be characterised as a latent variable problem. Suppose we had some data (e.g. labour force survey) that gave some indication as to the relationship, but in generally will not have a fully dataset integrated with main administrative data



A challenging issue to cover off properly

- **Could simulate job status as a function of welfare status (i.e. reverse direction of the red arrows).** A little unsatisfactory as causality-type questions are harder to explore
- **Existing models of work and welfare are hard to integrate.** Ideally need compatible setups in terms of time-step, populations, time periods, demographic splits etc.
- **Full Bayesian models can be quite slow, both to fit and project.** The time chronology is a bit unclear too – if someone remains off benefit, maybe this increases the probability they had a job in earlier quarters?

There is a broader question (not covered elsewhere) about how we make best use of other data/models that live outside the main administrative data.

2. Latent variables

Potential solutions

Could a more sophisticated model setup be derived that:

- Allowed for a probabilistic latent variable that evolved over time (in a manner consistent with the other observed variables)
- Leveraged external data on the relationship between latent and observed variables
- Runs efficiently.

Kalman filter type setups promising, but there's a fair amount to work through, including how it plays out in a long-term projection.

Unsure if Bayesian network theory can be applied easily or effectively

3

Component model structures

3. Component models

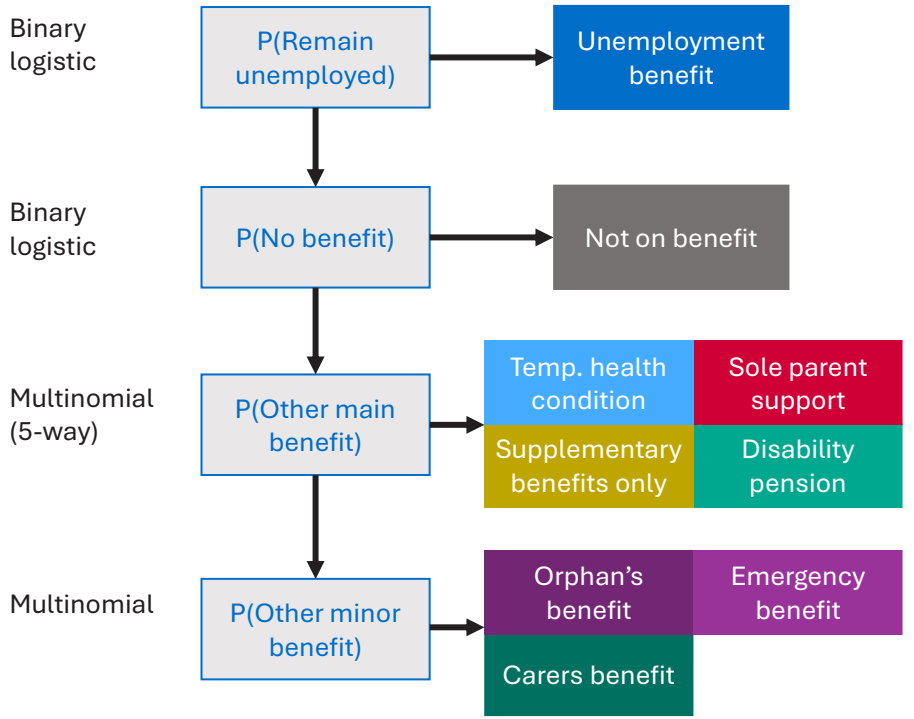
The problem

We have a lot of component models (dozens – or even >100). We also have choice on the exact time windows and model setup.

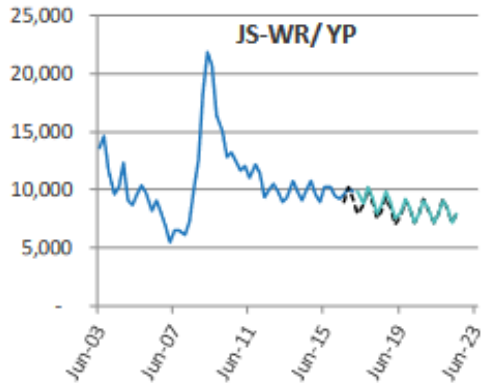
- There are many considerations around:
  - Speed to fit
  - Speed to score/apply
  - Model accuracy
  - Model stability over time
  - Smooth effects
  - Hypothesis testing
  - Handling time trends and incorporating economic factors
  - Control over factors and compatibility between sub-models

  Considerations may vary depending on relative importance of an individual model

Example of multiple component models for benefit state, NZ Welfare work



NZ benefit entries @2017, for main unemployment supports



3. Component models

Best approach far from settled

Fair variety in existing models

- Tailored regressions (GLMs / multinomials) fit to longer time windows

  Quite time consuming, more judgement required

- Fixed, simplified regressions fit to narrow time windows

  Tend to ignore trends and economic effects, miss key interactions

- Black box models (boosted decision trees or neural networks)

  Struggle with large disruptions (COVID), trends, economic effects, score speed

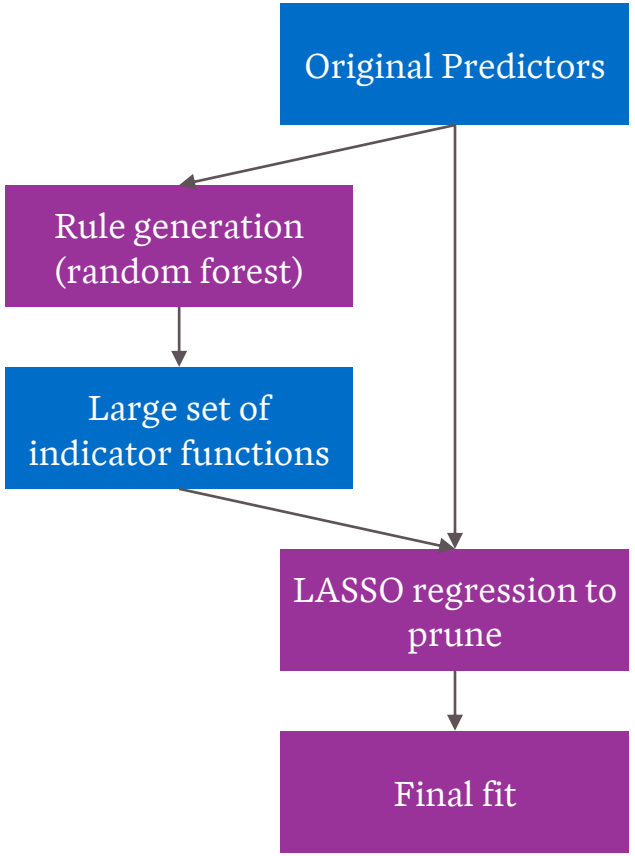
Most models have gravitated towards economic factors being a separate estimate (bolt-on or offset) – making the process less streamlined.

3. Component models

Potential solutions

- No real thoughts on the solution to economic sensitivities – two step modelling possibly unavoidable.
- Some algorithms out there (e.g. Rulefit) appear to offer black-box performance with a more streamlined scoring formula.
- Transfer learning ideas, or other types of representations, might allow models to 'share' variable generation steps to increase projection efficiency.
- Refits that start with, or leverage, the previous years model probably have merit.

Rulefit schematic



4

Dealing with dimensionality

4. Dealing with dimensionality

The problem

We can quickly get to a lot of 'predictor' variables, especially when you realise there are multiple forms (e.g. % of past year in prison vs % of past 3 years)

- More steps in the projection cycle → more risk of skews
- More complex model → Total parameters in a set of GLMs grows parabolically . Harder to keep all up to date
- Speed implications (fitting and projection)
- More correlations to check (parabolic) in projection outputs

Example - main NZ model drivers

- **Demographics:** Age, gender, ethnicity, education
- **Region:** MSD service region, TLA
- **Benefit receipt:** Current status and duration, historical receipt, benefit sanctions
- Parents' benefit receipt
- **Public housing:** Status, Rent subsidy, Income, SAS need scores
- Type of health condition
- **Household:** Partner status, children (age and number). For PH, additional household vars like number of bedrooms needed.
- **Other government service history:**
  - Corrections (e.g. recent time in prison)
  - Youth Justice (e.g. # youth offences)
  - Child protection history (e.g. time in placements)
  - Education (e.g. highest attainment)
- **Macroeconomic:** Regional unemployment rate, market rents, AWE

4. Dealing with dimensionality

Potential solutions

- Machine learning for all models, reduced human intervention. Requires a very strong validation framework.
- Non-parametric methods that by design will keep plausible relationships between variables. Just as the bootstrap allows plausible simulation in some contexts, approaches could be tailored to this type of simulation.

5

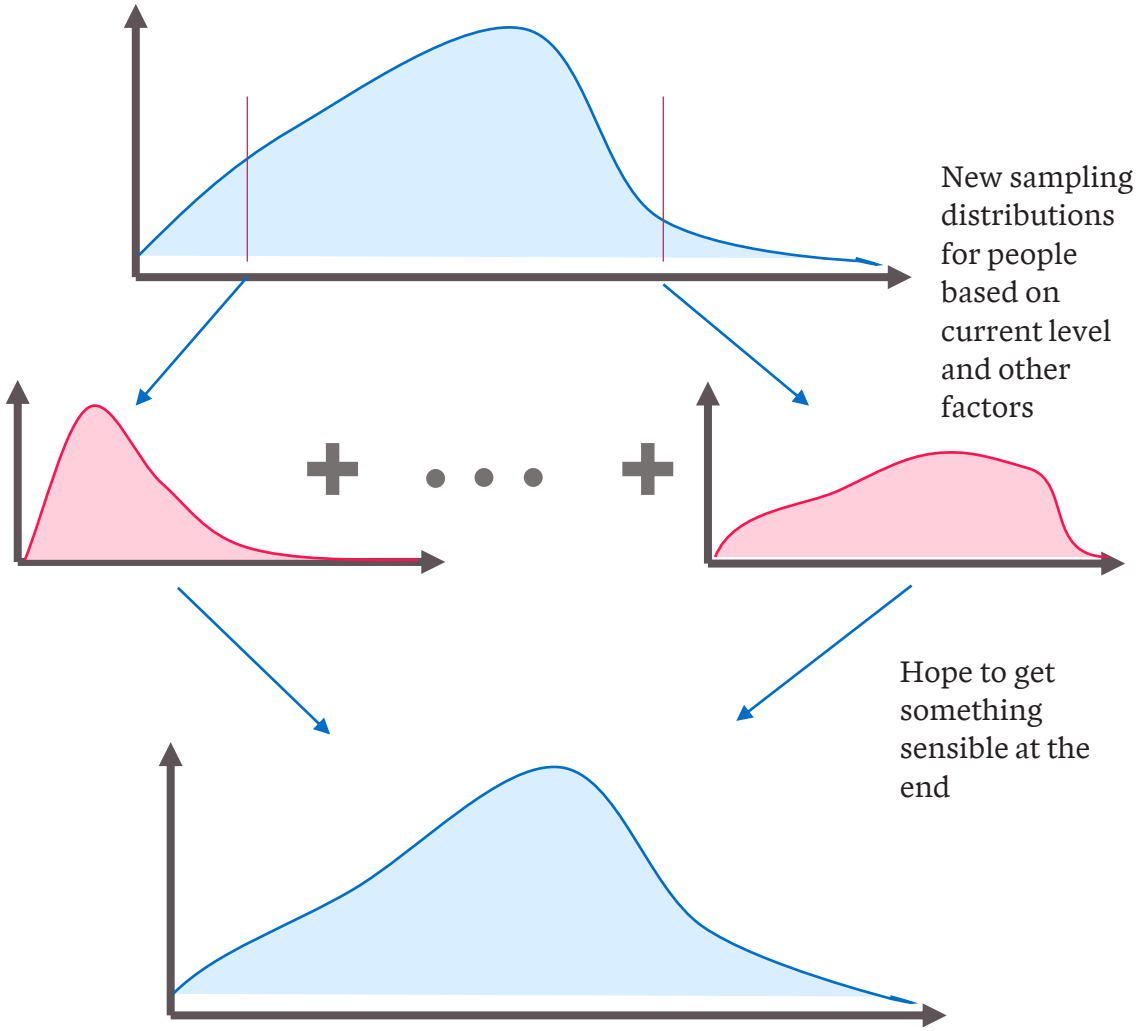
Continuous variables

5. Continuous variables

The problem

Categorical variables are ‘easy’ to model – define the categories and model transition probabilities and unleash standard modelling techniques.

Continuous variables (e.g. a person’s income) is much harder. The main issue is that modelling is distributional – we are not interested in the mean income, but the distribution of income for someone given their current point on the distribution curve.



5. Continuous variables

Potential solutions

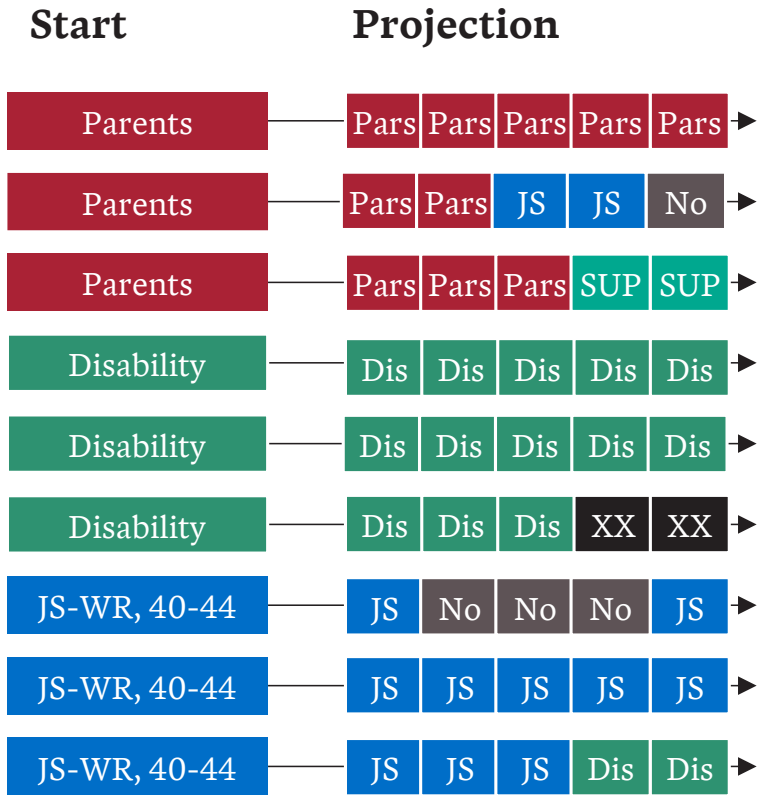
- Convert continuous into a categorical (e.g. 5 income levels) – quite common, although can throw a fair bit of insight away.
- Distributional regression models (e.g. Beta-regressions) do exist, but less routine, and often a lot of checking needed.
  - Robust methods are needed for real-world distributions.
  - A setup with good memory properties ideal (e.g. could give weight to different values in the previous X years, as well as other predictors)

6

Validation

6. Projection validation

- Modelling is best with a simple objective to maximise (e.g. AIC, loss, gains)
- But this is hard to do for the overall model – how should we be validating?
  - Outputs are a set of timesteps with status/variables in each
  - Typically validation is lots of checking on trend for different cohorts, comparing to historical (or other expected)
  - Back-testing can be useful, but even then, it's a lot of summary comparisons. And significant practical difficulties
- Actual vs expected measures (e.g. misclassification rate at time *t*) harder to define for later timesteps, since the model will have a different starting point at time *t*-1.
- Some distributional checks are possible. However much harder to say something about pathway plausibly.
- So if we were running a 5-year back-test, and most interested in welfare state over time, is there a 'best' single measure for validation, covering both pathways and final distribution?



Is this a 'good' projection?

7

Model size & speed

7. Model size and speed

The problem

Data volumes are large (e.g. a ‘full’ simulation output >100gb) , and the projection stage is usually measured in hours (potentially days).

Sheer size of datasets can materially slow down the modelling and projection process.

Potential solutions

Much of it solved. On the projection side:

- Aiming for ‘low code’ in model scoring
- Program language-specific enhancements
- CPU Paralellisation
- More adaptive or flexible dataset sizes. For example, starting with ‘bins’ of people with similar characteristics
- Selective in outputs.

But, probably more to be done:

- Smarter dataset structures out there, particularly for non-independent projections (e.g. linked households)
- GPU computing probably better, but requires a pretty particular skillset and commitment

8

Differential data windows

8. Differences in data time windows

The problem

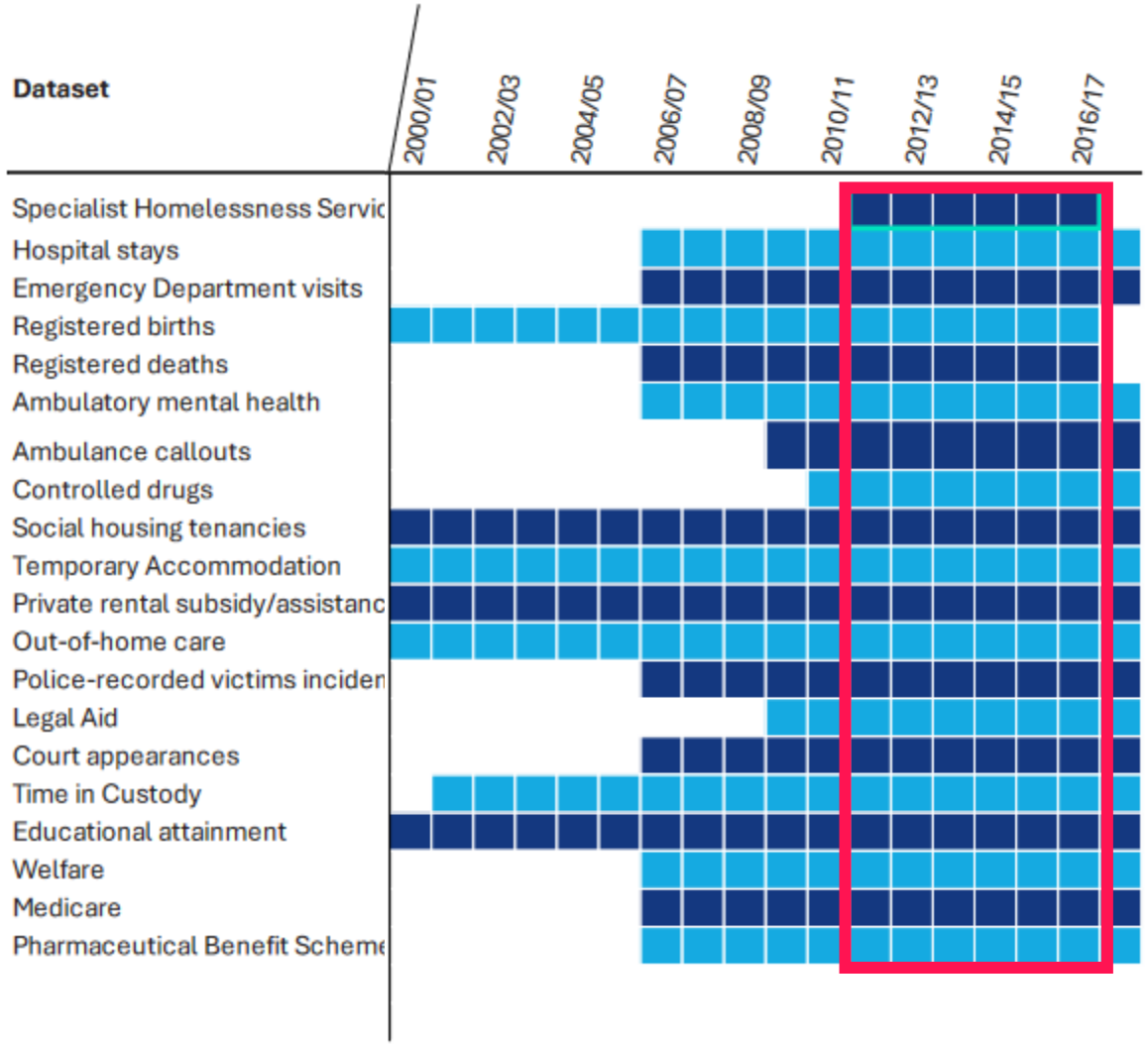
For linkage projects, datasets will not line up. In the example, we only have 6 years common to all – if we want a 'Homelessness in the past 3 years', that then reduces to 3 years of usable data, which feels potentially wasteful.

- Particularly an issue for methods using longer time windows (e.g. pathway sampling) – in some ways the one timestep transition approach solves much of the issue
- More generally, the length of data reduces ability to model (e.g. longer history variables, intergenerational effects)

Potential solutions

- Smarter use of imputed and missing variables
- Back-casting before forecasting

A good linkage result – 6 years of common data, but still potentially ignoring a lot



TAYLOR FRY

Sydney
Level 22
45 Clarence Street
Sydney NSW 2000

www.taylorfry.com.au

Hugh Miller
Principal, Taylor Fry
hugh.miller@taylorfry.com.au
02 9249 2926