A sensitivity analysis of the birth cohort model for tertiary education

Anke Weber, Andrea Saltelli and William Becker

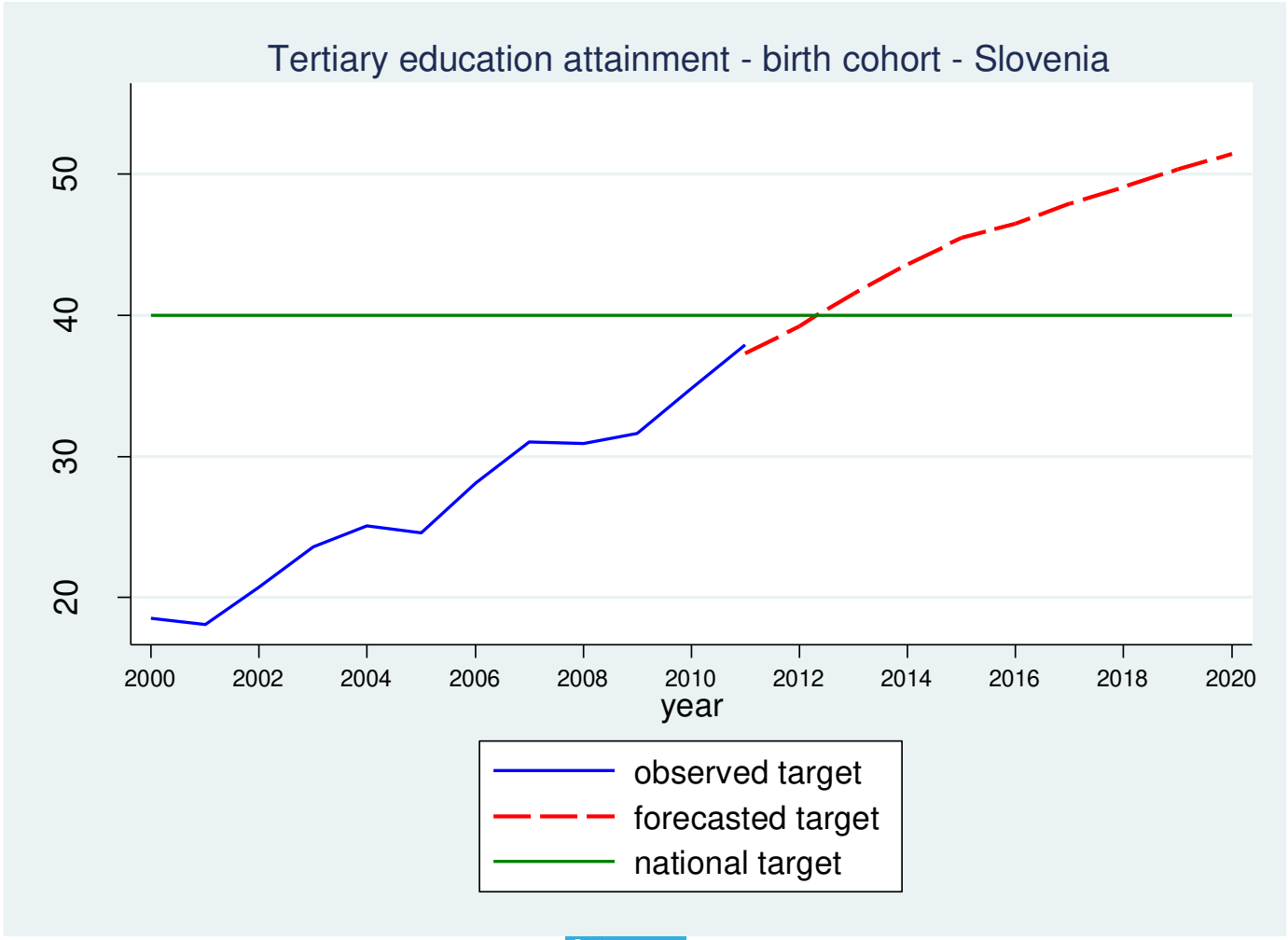
Nice, 1-4 July 2013

Background

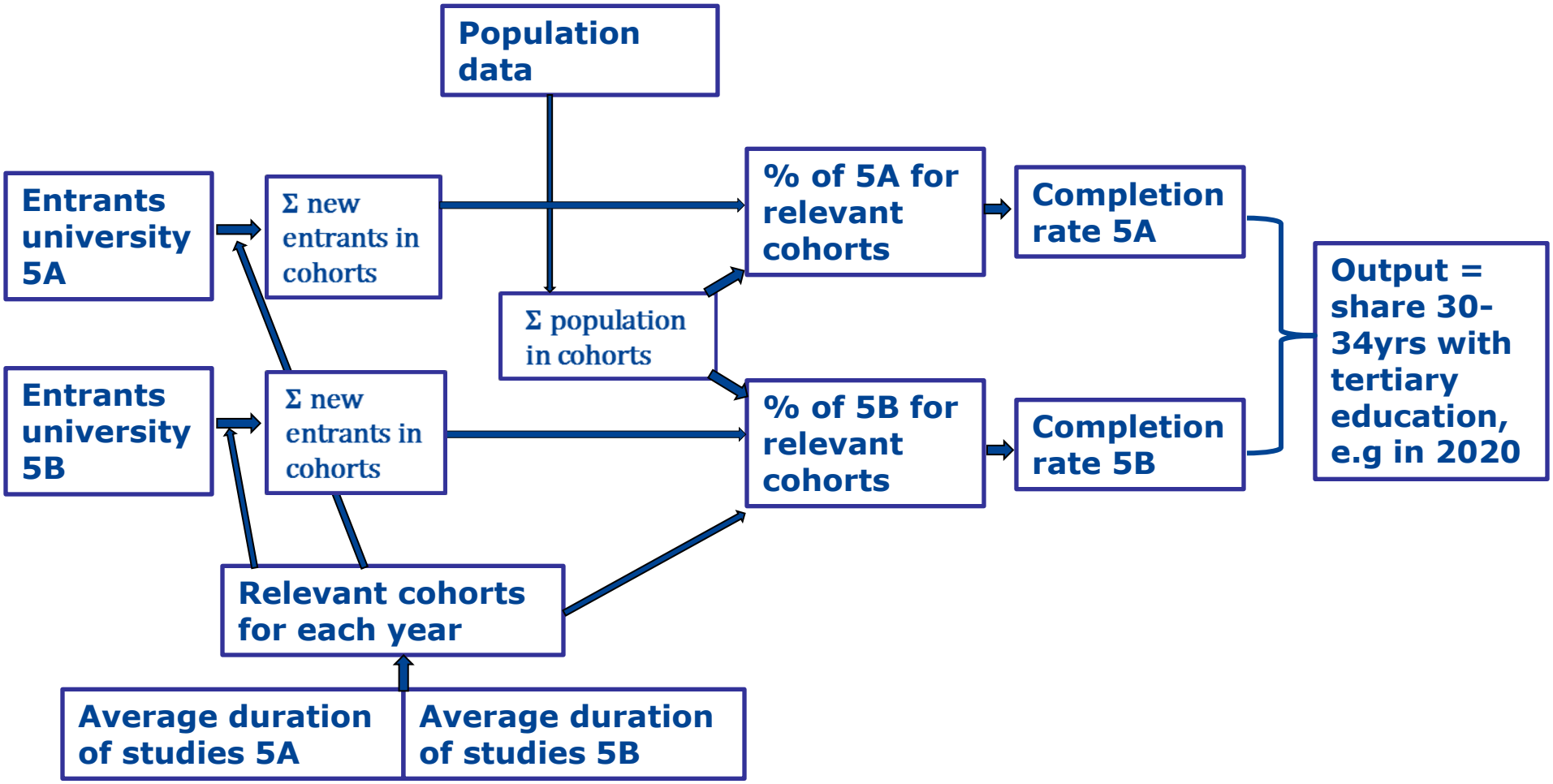
- **JRC developed birth cohort model to monitor whether countries achieve 2020 target on tertiary education benchmark**
- **Benchmark = Share of 30-34 years old that have completed tertiary education**
- **Use birth cohort model: data on new entrants to tertiary education (ISCED 5A, 5B) that have the right age**
- **Uncertainty?**
  - **Extrapolation of missing data**
  - **Underlying assumptions on average duration of studies and completion rate**

Model output

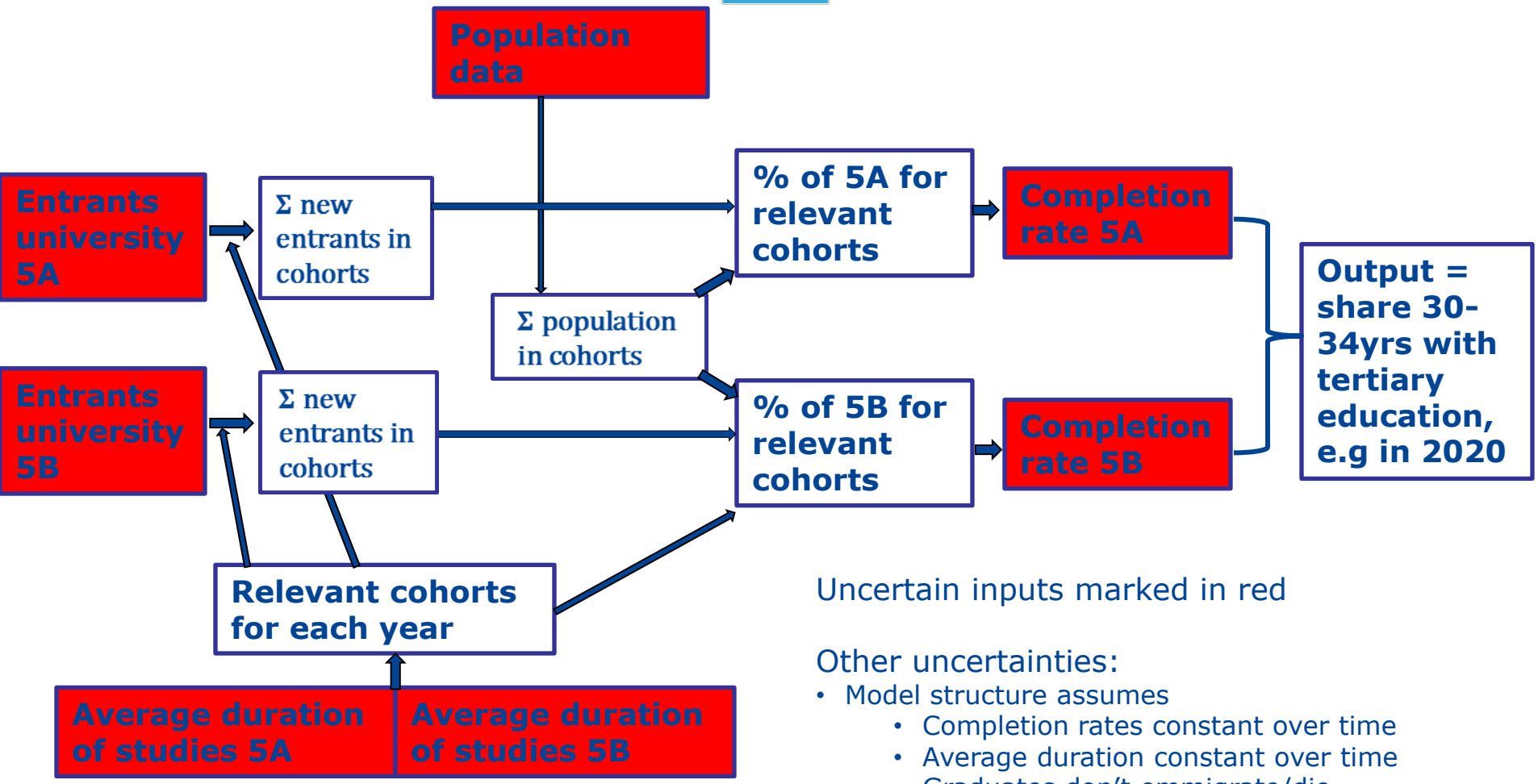
Example for Slovenia



Model overview



Model overview – uncertainties

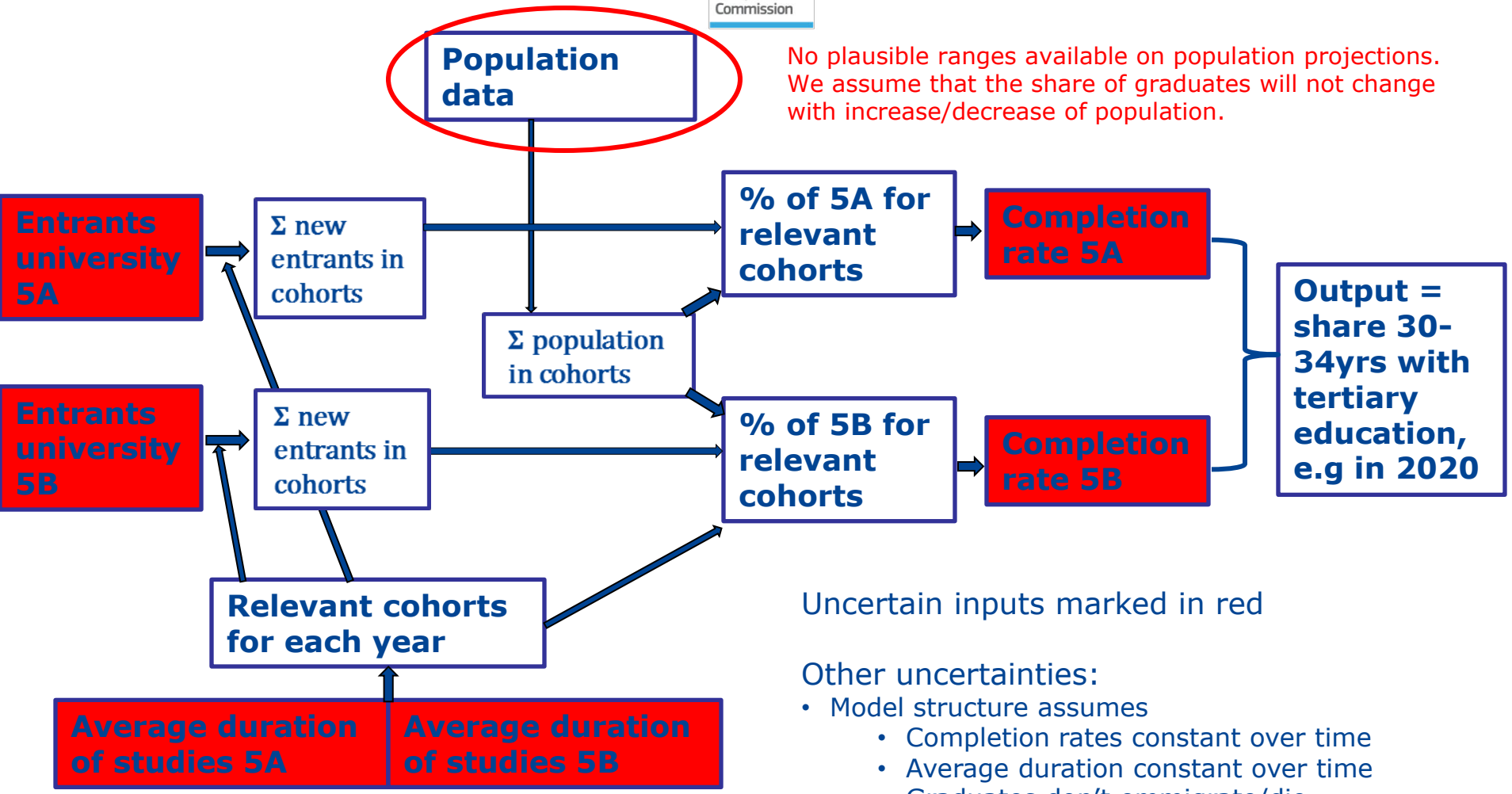


Uncertain inputs marked in red

Other uncertainties:

- Model structure assumes
  - Completion rates constant over time
  - Average duration constant over time
  - Graduates don't emmigrate/die
  - Graduates don't immigrate

Model overview – uncertainties



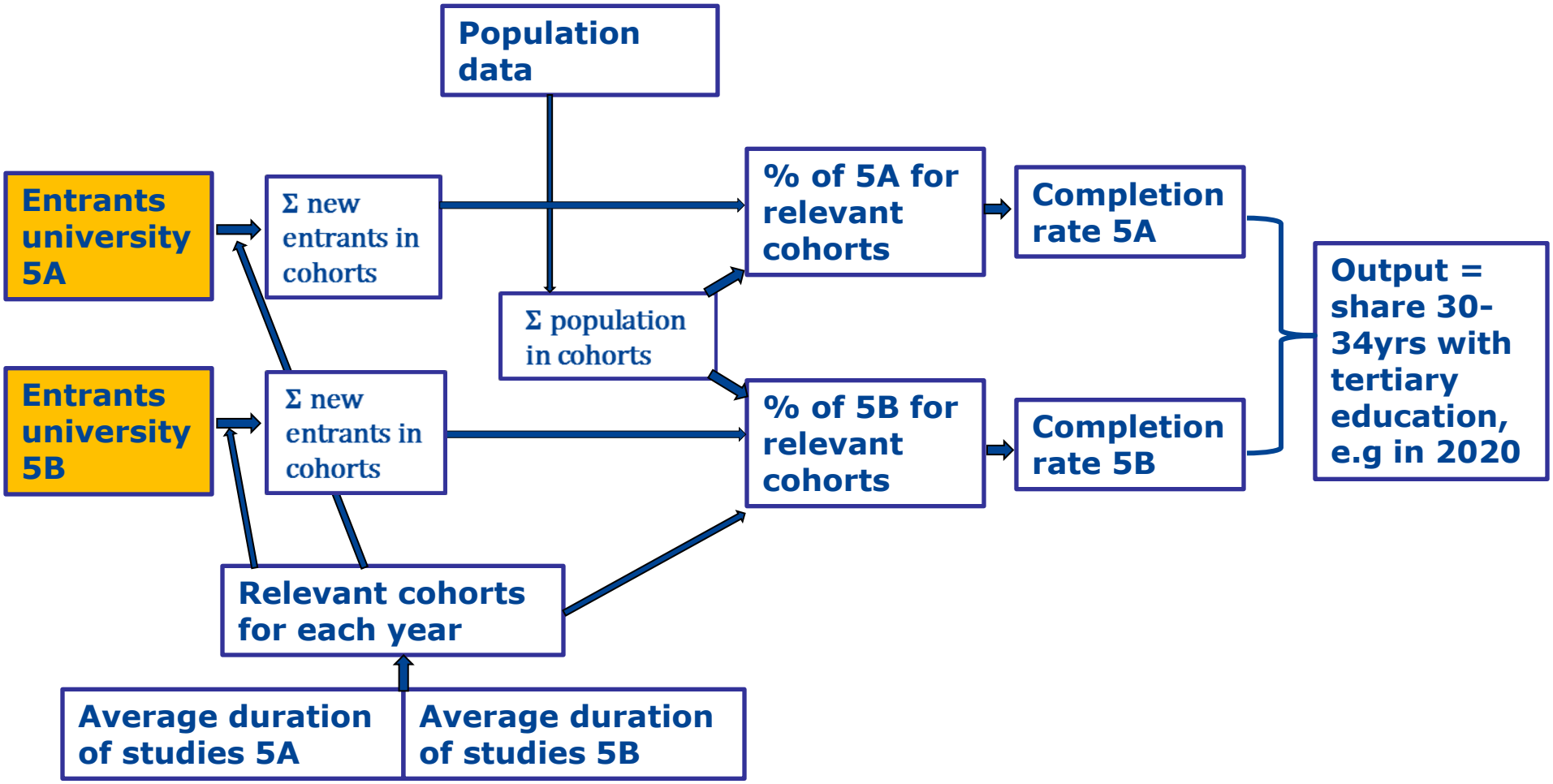
Uncertain inputs marked in red

Other uncertainties:

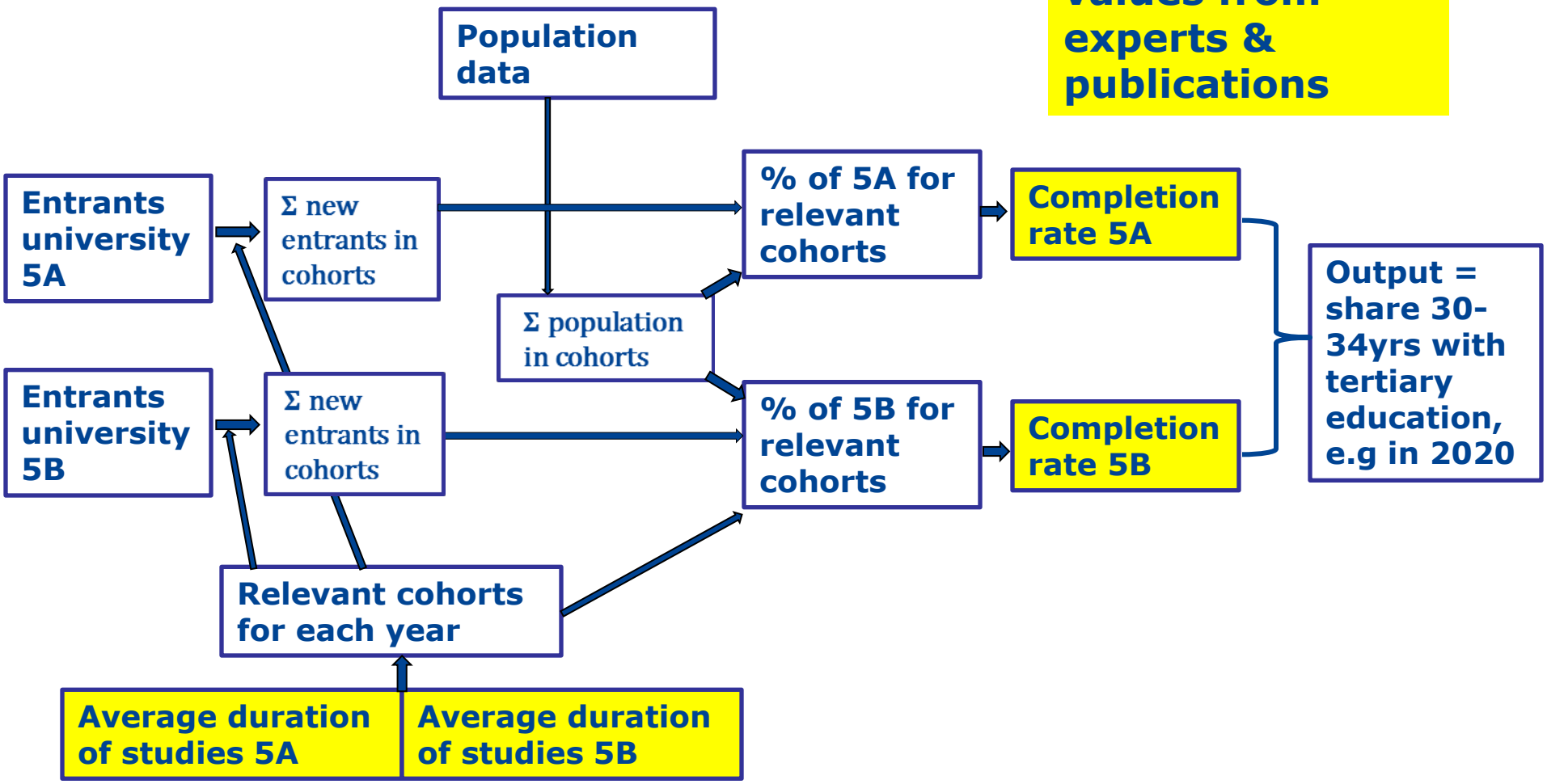
- Model structure assumes
  - Completion rates constant over time
  - Average duration constant over time
  - Graduates don't emmigrate/die
  - Graduates don't immigrate

Random variables

Extrapolation



Random variables



Black box summary

Inputs:

- **Entrants 5A, 5B**
- **Average duration 5A,5B**
- **Completion rate 5A/5B**

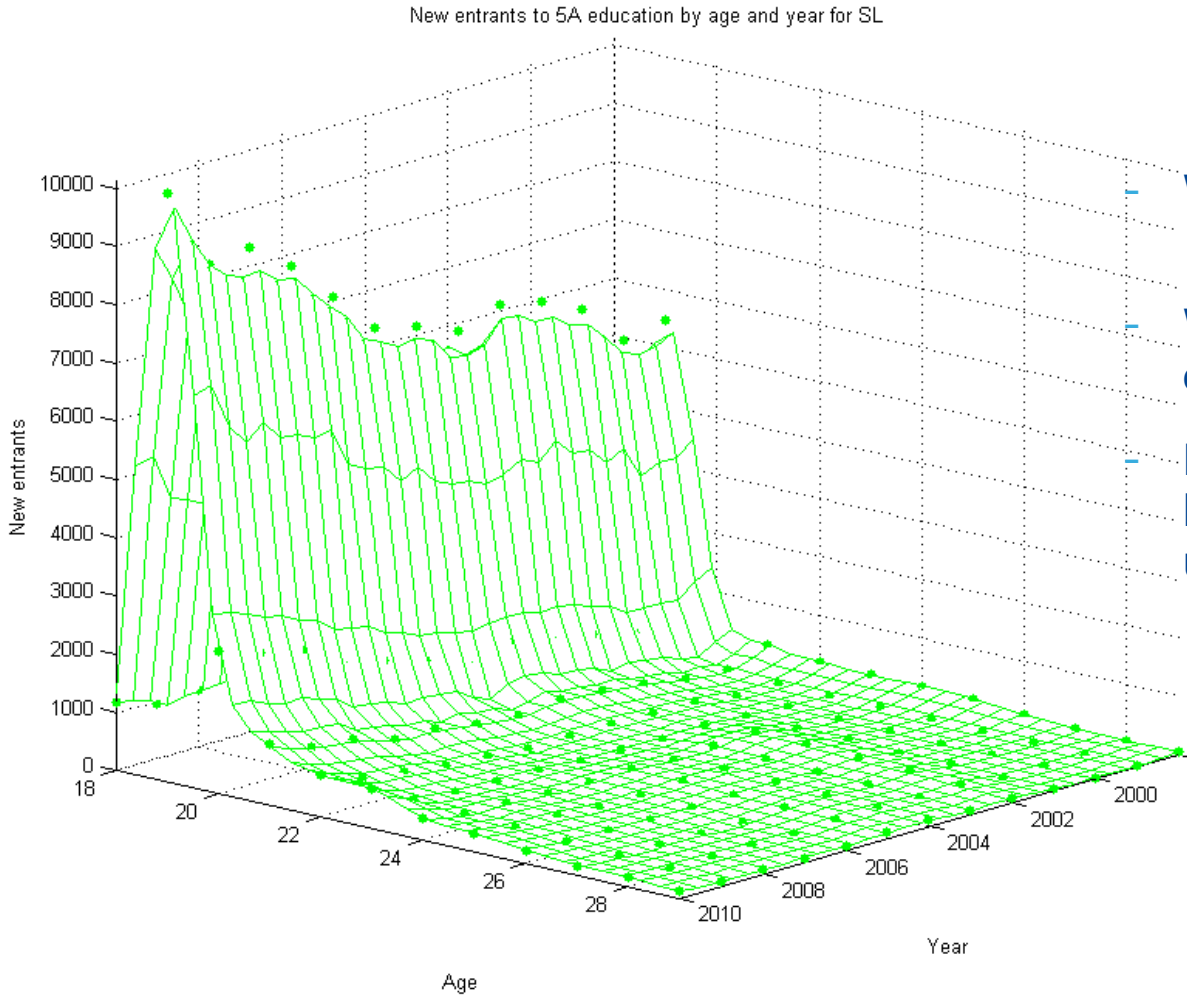
Model

Output = share 30-34yrs with tertiary education, e.g in 2020

- Average duration and completion rates are easy to implement into a SA, being continuous parameters with ranges
- Data sets for 5A/5B entrants are however incomplete, and we need to account for the uncertainty here due to extrapolating this data.

Extrapolation

A large source of uncertainty is due to missing data points in the entrant data, which must be extrapolated.



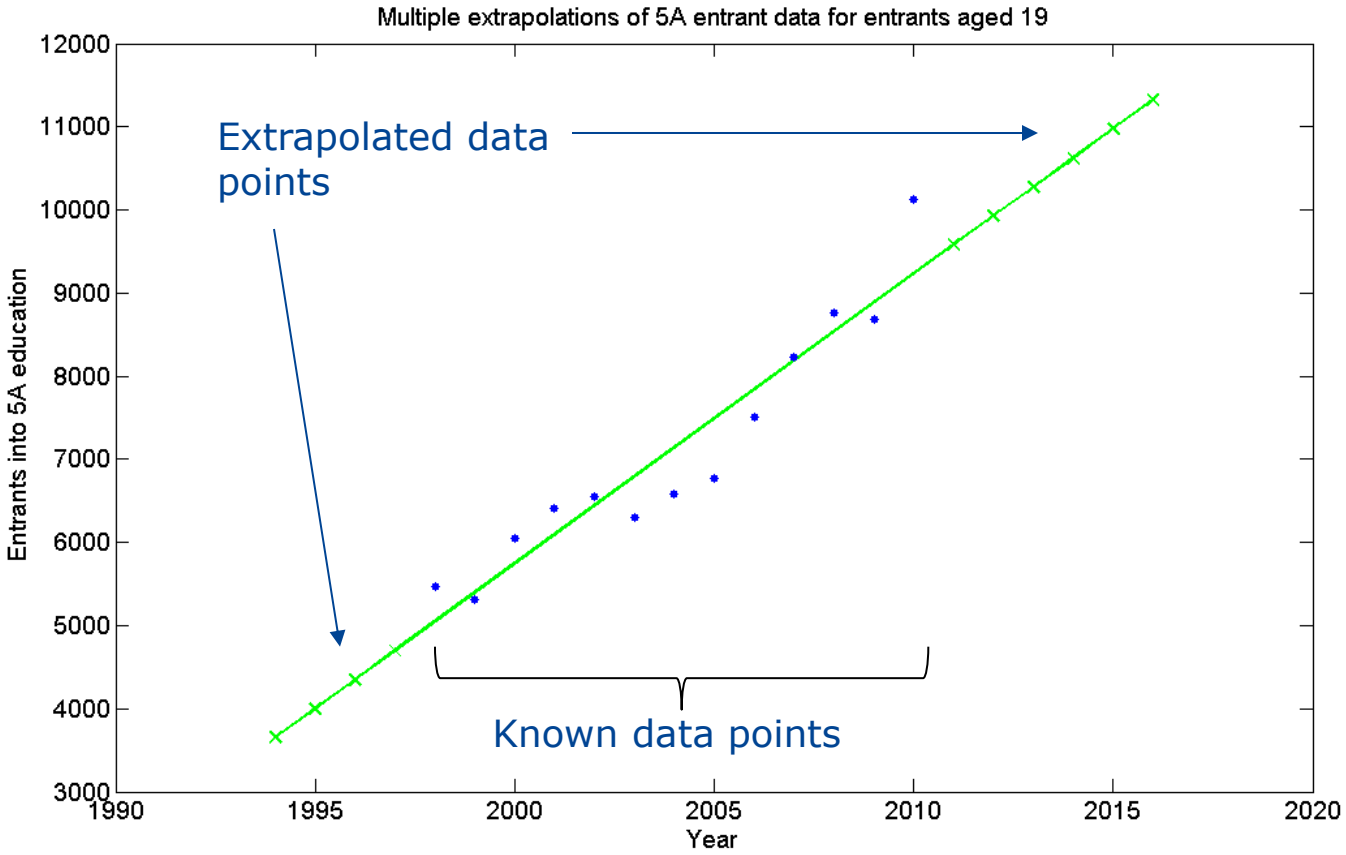
We require data for years 1994-2016

We are forced to extrapolate

But this is very uncertain… how to account for this uncertainty?

Extrapolation

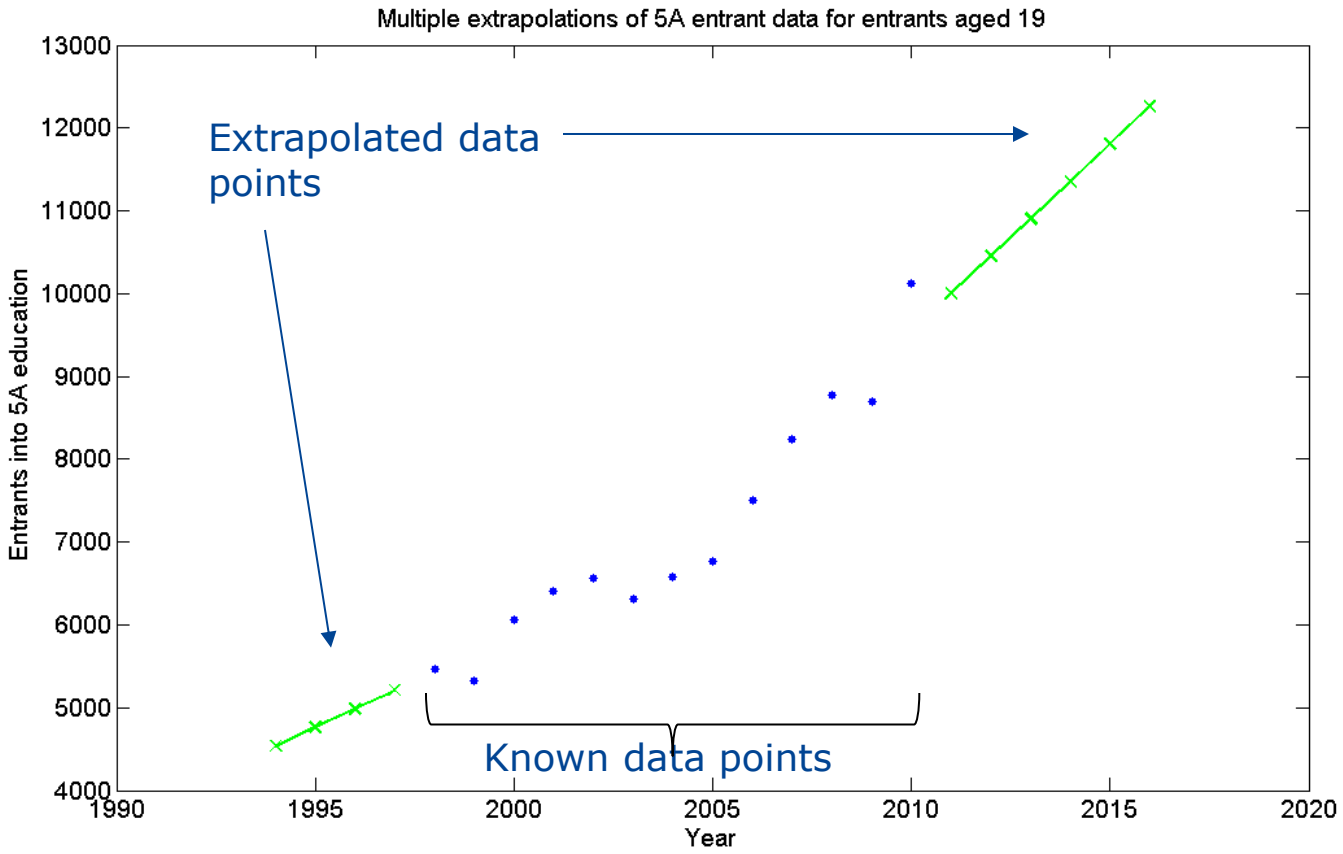
- We cannot justify assuming any trend more complex than linear
- It is unknown whether the extrapolation should follow only the closest points or use the entire data set



Here we can try using all the data points to extrapolate….

Extrapolation

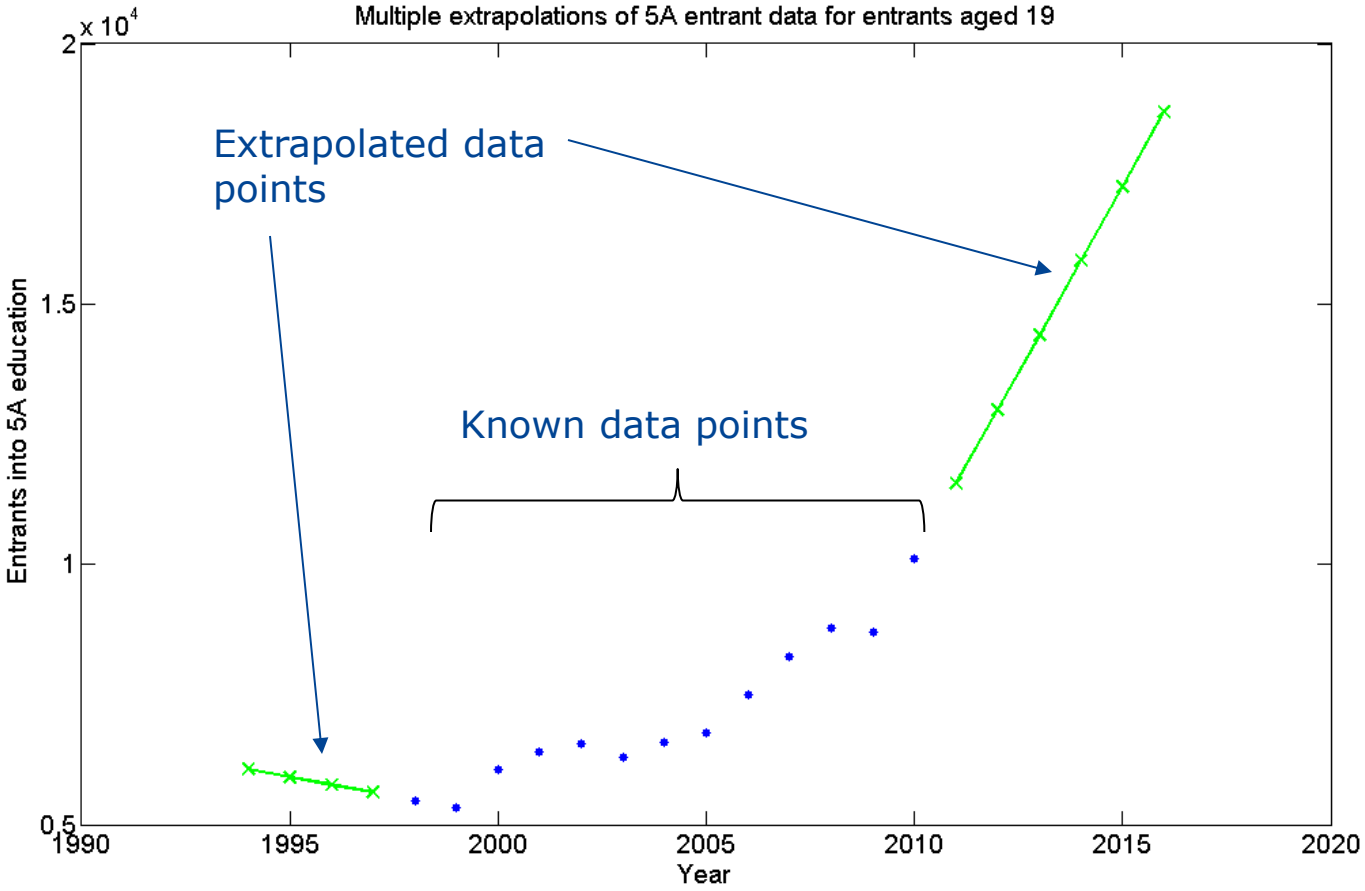
- We cannot justify assuming any trend more complex than linear
- It is unknown whether the extrapolation should follow only the closest points or use the entire data set



Here we exclude the furthest five points – note trend is more local

Extrapolation

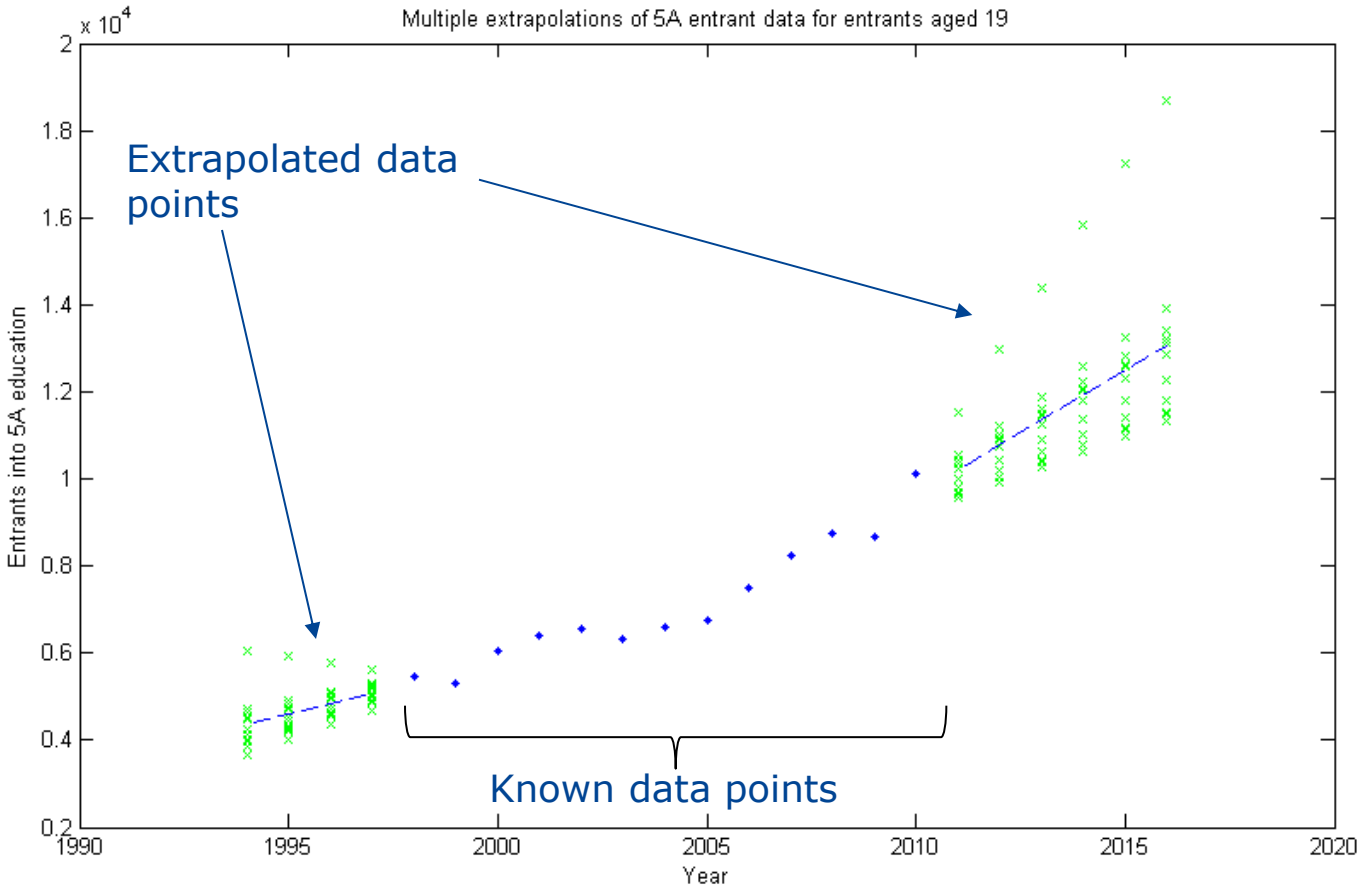
- We cannot justify assuming any trend more complex than linear
- It is unknown whether the extrapolation should follow only the closest points or use the entire data set



Here we use only the closest two points – very local trend. Not very likely but within the realms of possibility!

Extrapolation

- We cannot justify assuming any trend more complex than linear
- It is unknown whether the extrapolation should follow only the closest points or use the entire data set



If we do this for all possible reduced sets of data points (by removing points in order), we have 12 possible extrapolations for the years before the data, and 12 more for the years after the data.

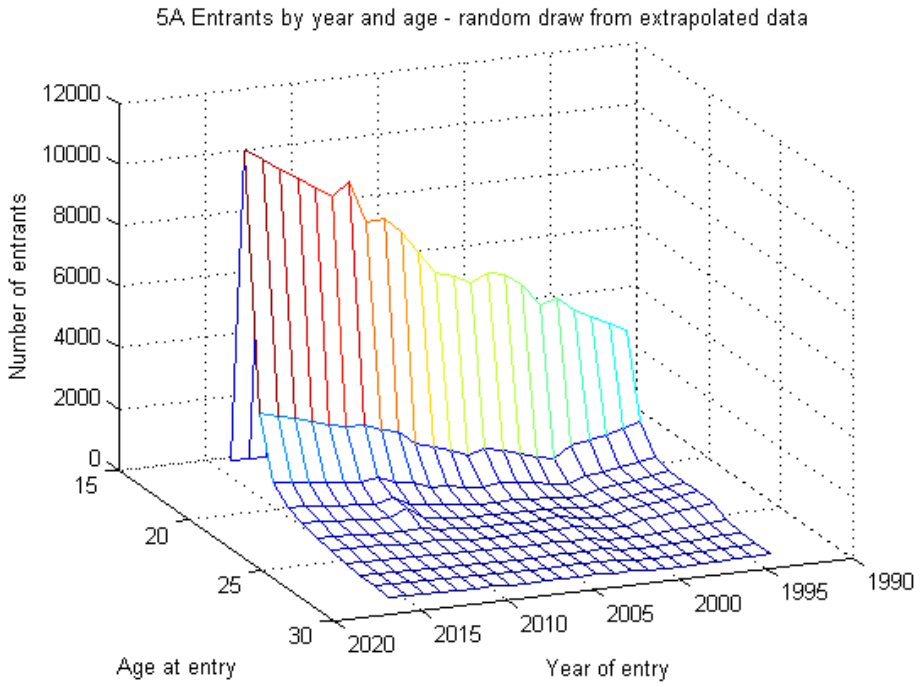
I.e. a total of 144 possible datasets which (in some way) express our uncertainty about extrapolating the data.

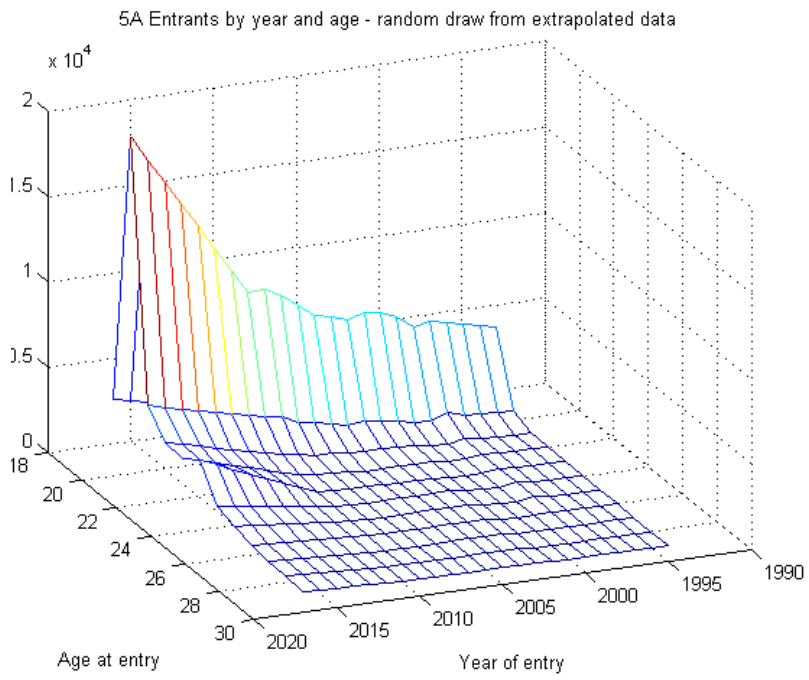
Extrapolation

- We cannot justify assuming any trend more complex than linear
- It is unknown whether the extrapolation should follow only the closest points or use the entire data set

---

- We extrapolate for each age group separately, but assume a "common trend", i.e. a given dataset draw will use the same number of points for every age group.





Uncertainty quantification (other parameters)

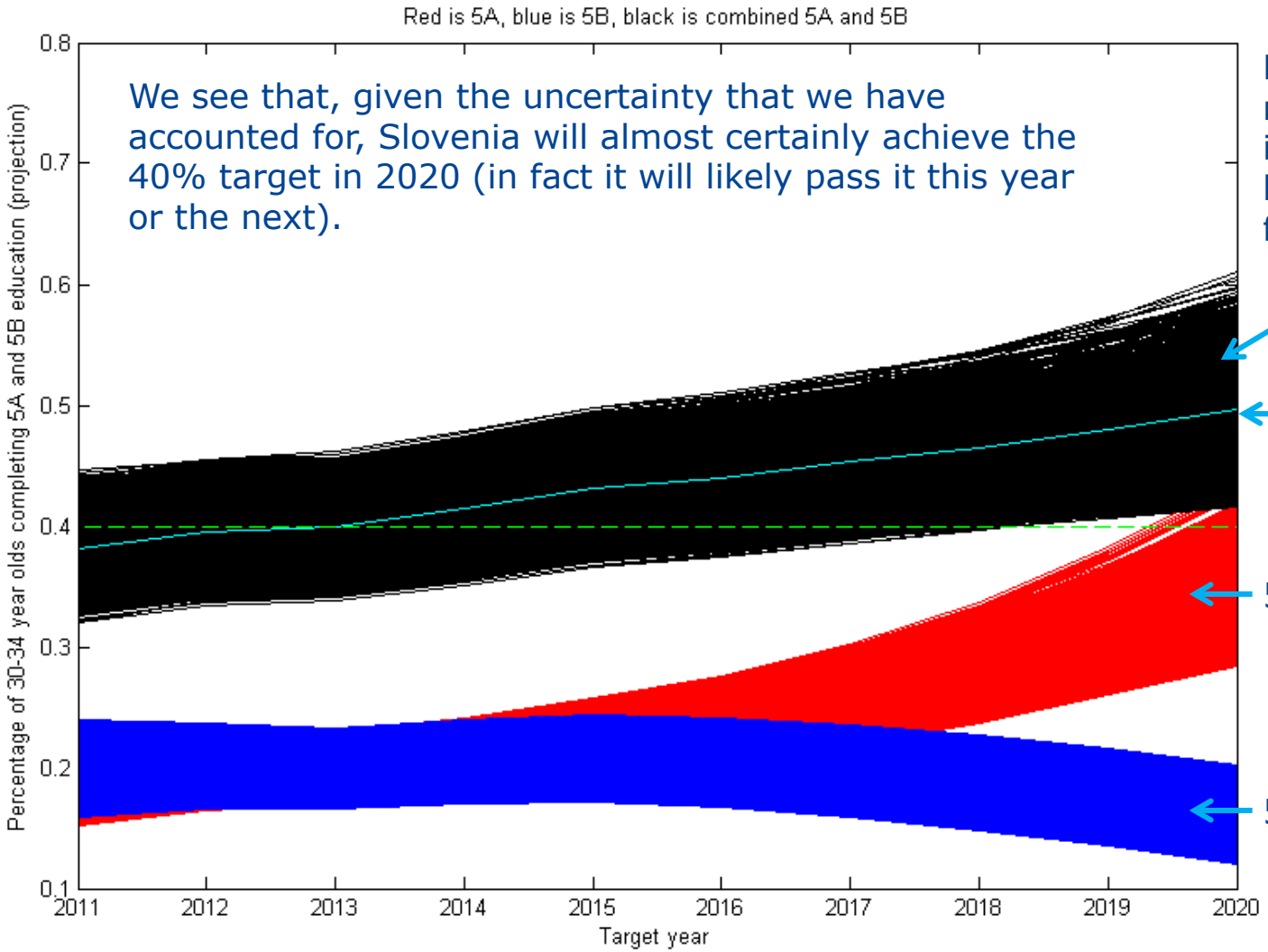
- Average duration of studies (5A/5B):
  - 1 data point from EU publication & expert opinion
  - Create random variable:
    - ISCED 5A → range: 3-5
    - ISCED 5B → range: 2-4
- Completion rate (5A/5B):
  - 2 data points (2008, 2011) from OECD & confirmed experts
  - Check completion rates of countries, which have high correlation with SI on tertiary education, i.e. UK, PL (corr >0.98) → assume reasonable variation lies within the range of SI-UK-PL
  - Create random variable:
    - ISCED 5A → range: 0.61-0.81
    - ISCED 5B → range: 0.53-0.74

Summary of Sensitivity Analysis

Input	Type	Min	Max
5A data (lower)	Discrete	1	12
5A data (upper)	Discrete	1	12
5B data (lower)	Discrete	1	12
5B data (upper)	Discrete	1	12
5A duration (yrs)	Continuous	3	5
5B duration (yrs)	Continuous	2	4
5A completion rate	Continuous	0.61	0.81
5B completion rate	Continuous	0.53	0.74

- Monte Carlo simulation using 1024 points per variable (a total of 1024*(2+8)=10240 model runs). Model is computationally cheap.
- Sobol' sequence used in conjunction with standard MC estimators to estimate first and total-order sensitivity indices
- All distributions assumed uniform and independent
- Output is target value in 2020

Results - UA



Results of 2048 random points in the input space; each line represents one forecast

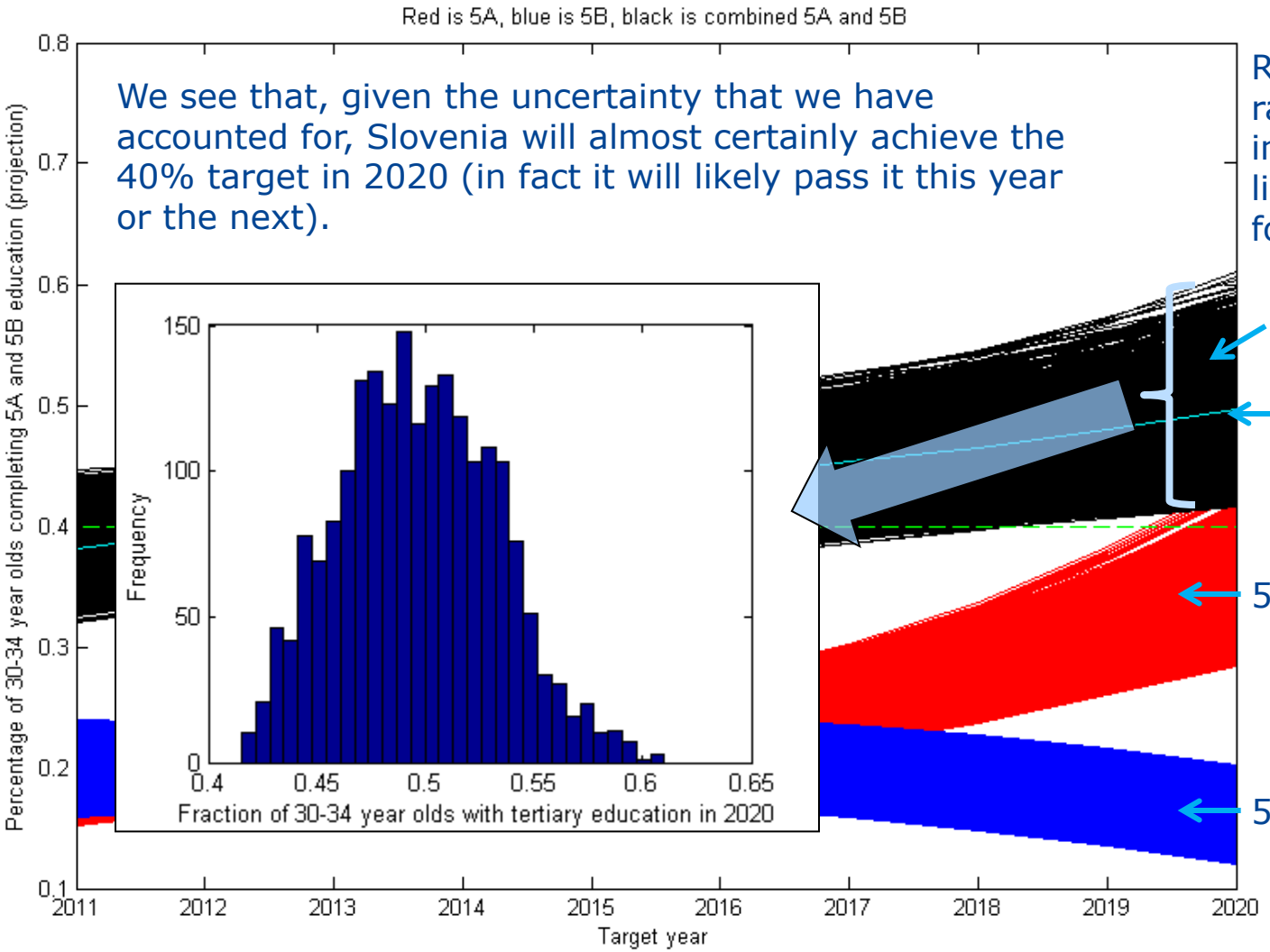
Combined 5A+5B (target)

Mean target value

5A

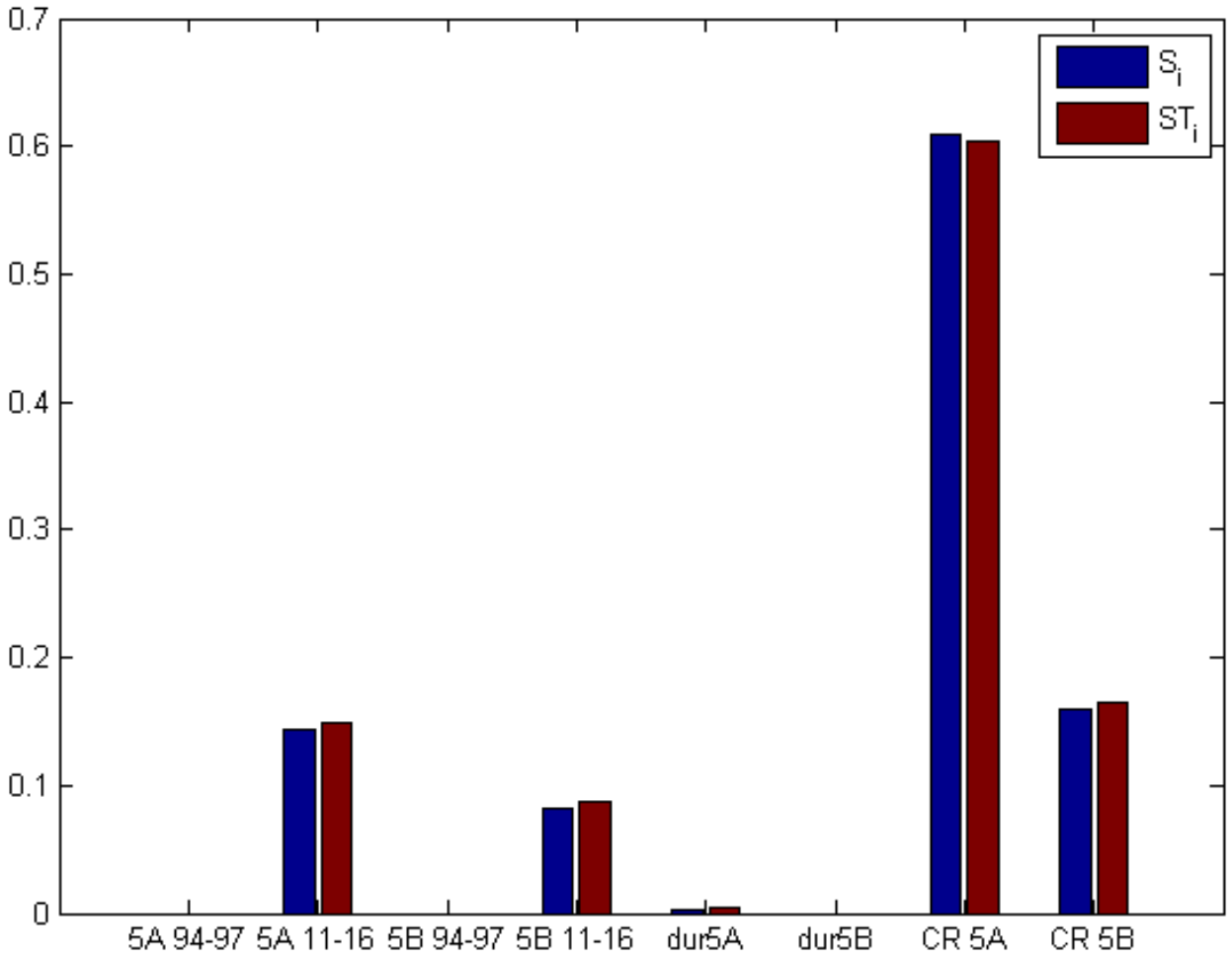
5B

Results - UA



Results of 2048 random points in the input space; each line represents one forecast

Results - SA



- We see that the most important variables are the completion rates
- The uncertainty in the data sets has only a small effect in comparison
- The duration of studies not important. This is because the large majority complete their studies in their early 20s, so changing duration by a few years will not affect the target much in 2020.

Open questions

How can we improve the estimate of output uncertainty when we cannot accurately quantify input uncertainty?

- Correct uncertainty for population data (realistic scenario – versus unrealistic scenario) ?
- Implement dynamic inputs, i.e. completion rate and average duration?
- Improvements to expressing uncertainty due to extrapolation?
- Improvements of information base for ranges of input variables, i.e. completion rate and average duration of studies? Expert elicitation.

Conclusions

- We have tried to perform an uncertainty analysis on a difficult "real-world" problem
- We have used a novel approach to deal with the uncertainty created by extrapolating data
- We find that it can be extremely difficult to quantify uncertainty accurately without under or overstating it. This is a serious problem for practitioners. This is caused simply by an absence of data.
- For the birth cohort model, we have discovered that Slovenia will very likely achieve the 2020 target.
- Future work:
  - More accurate elicitation of input uncertainty
  - Running SA for all EU countries