Statistics Flanders Research Seminar

Estimation Models for the Number of Immigrants with Observational Delay

Master's Programe in Statistics and Data Science
European Master of Official Statistics (EMOS)

Lie HONG
2024-May-21

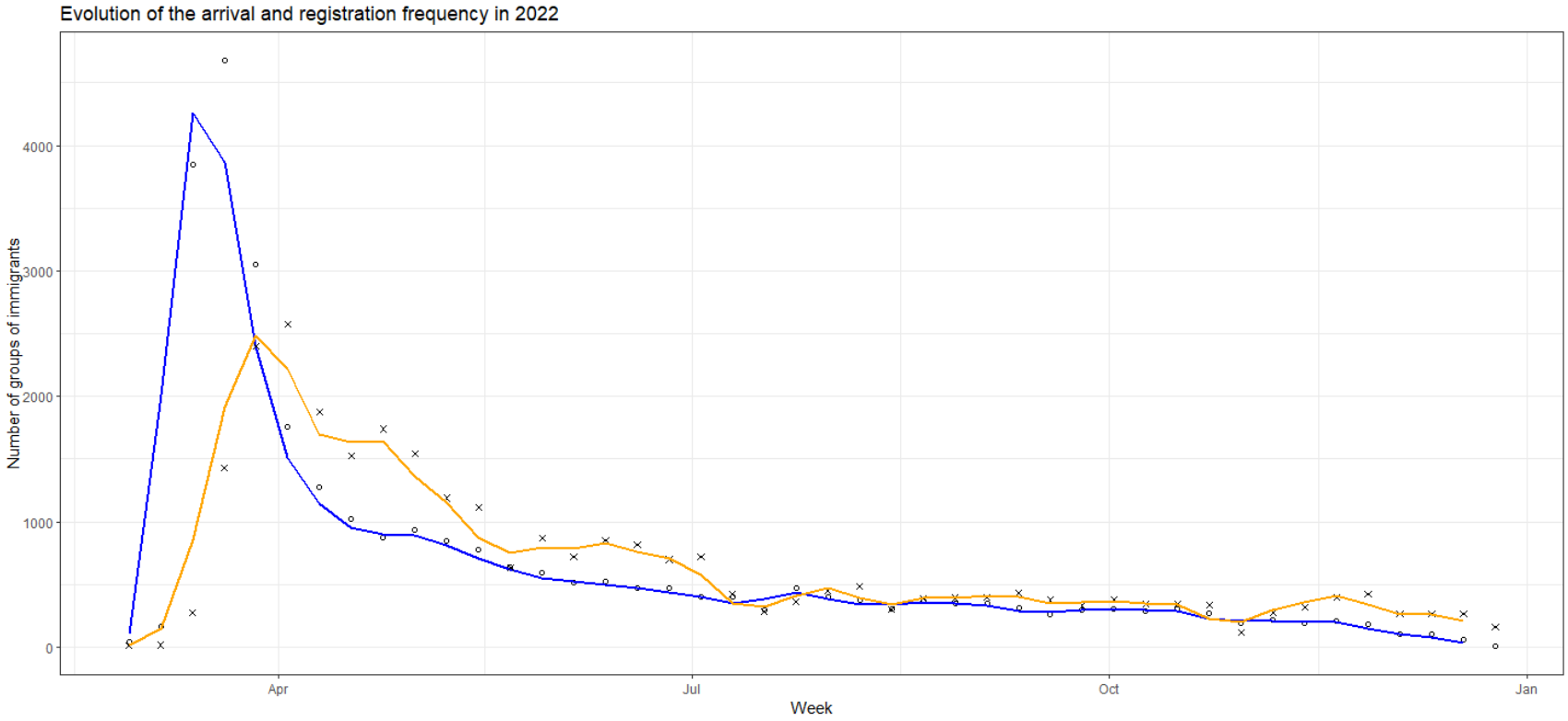
Supervised by Prof. Dr. Katrien Antonio
& Dr. Jorre Vannieuwenhuyze

Contents

(1) Overview of our project

1-1) Stating phenomena

Russo-Ukrainian War triggered immigrants, since 2022-Feb-24



(1) Overview of our project

1-2) Research goals

- Estimate the weekly and daily arrivals of immigrants
- Evaluate the methods applied to our data
- Unravel the factors influencing administrative delays

1-3) Challenges to overcome

- Big data like structure – volume, complexity, errors, …
- From data preprocessing to model selection
- Applicability – robustness, flexibility, functionalities, nowcasting …

(2) Data and methods

2-1) Data properties

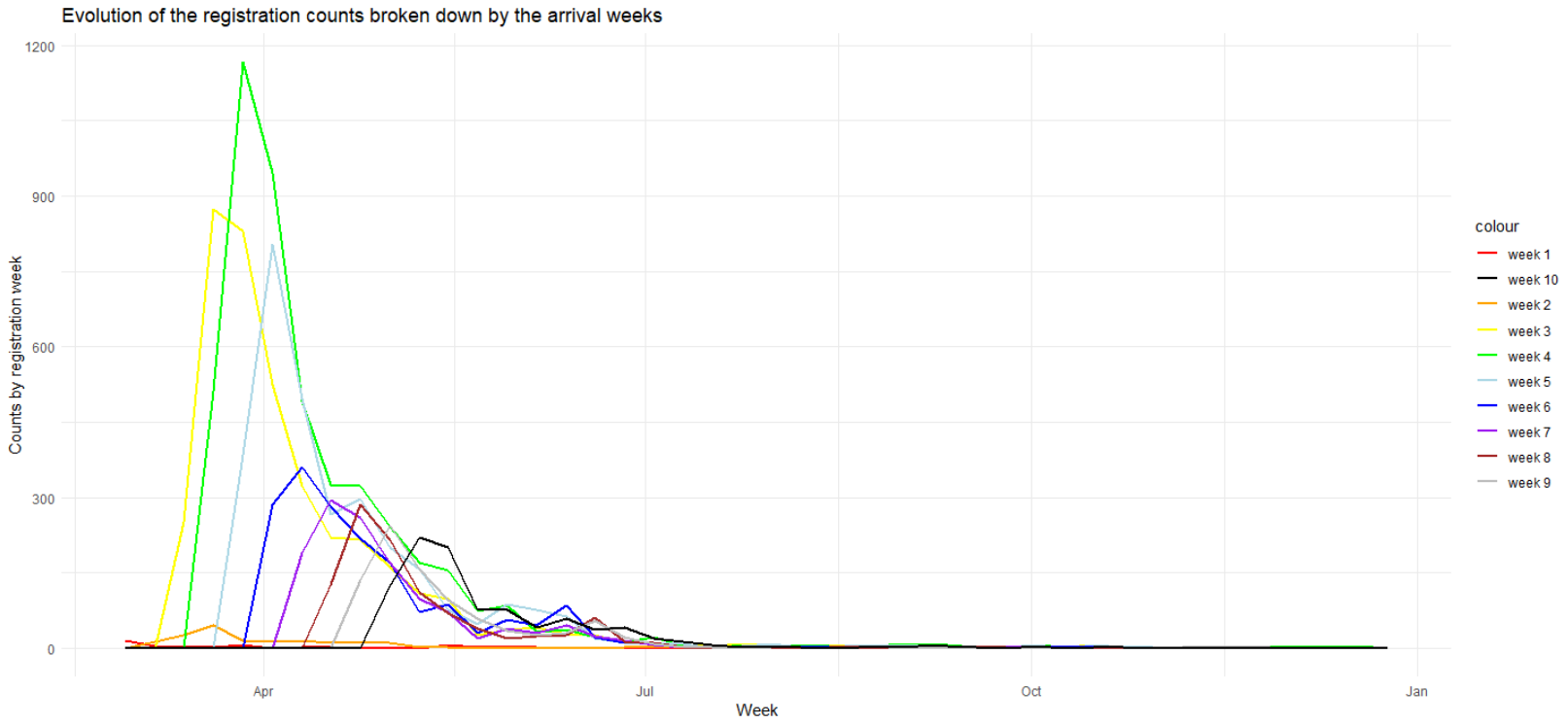
- complex – 38881 groups, comprised of 69936 immigrants
- 95.5 % delayed registration
- 3.813 weeks of delay on average
- registration: every Saturday

2-2) Methods

- Main reference – Verbelen R. et al. (2022)
- Borrow wisdom from actuarial science
- Chain Ladder: an industrial standard
- Mack's non-parametric approach & Poisson and negative binomial

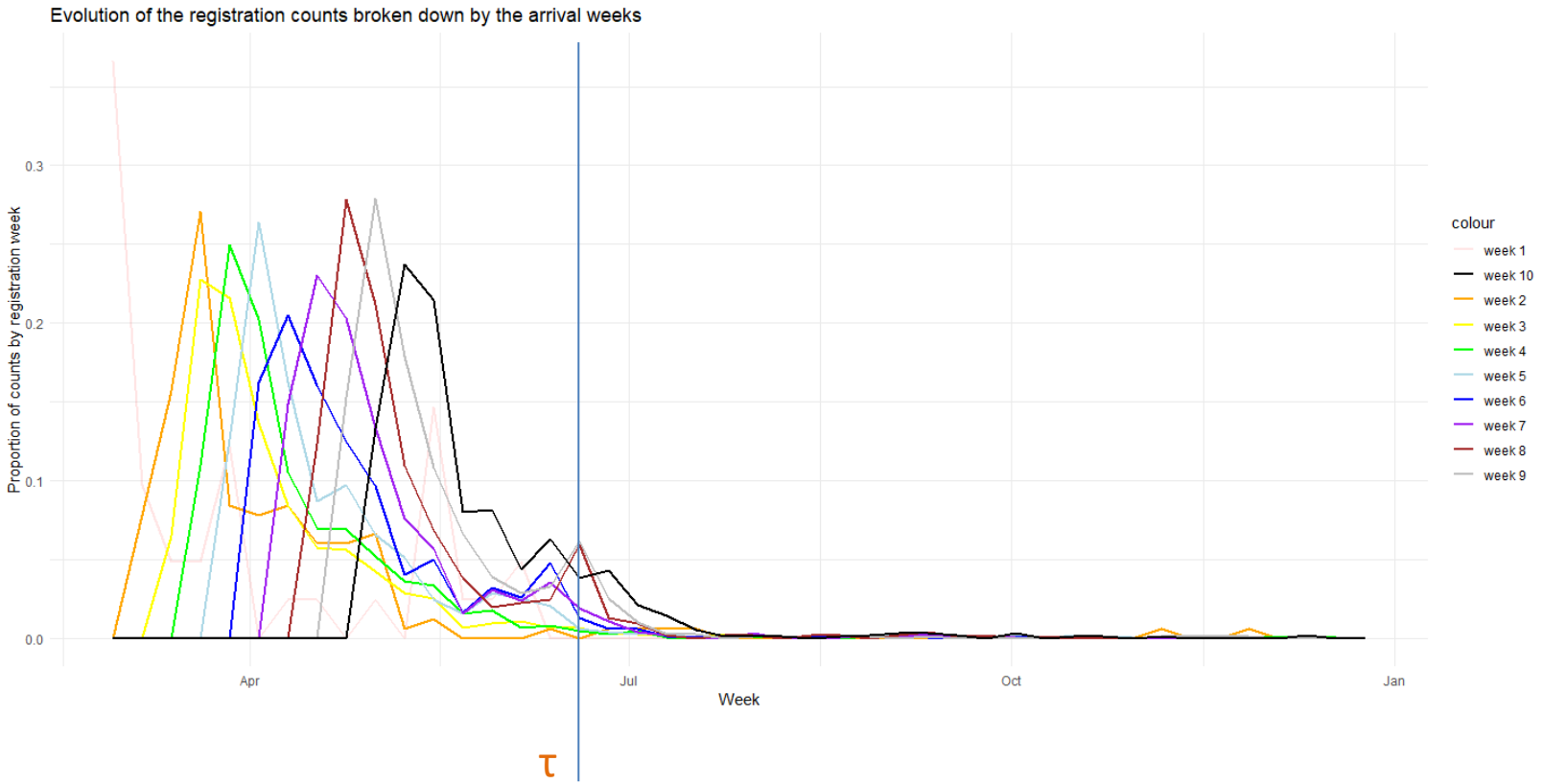
(2) Data and methods

2-1) Data properties



(2) Data and methods

2-1) Data properties



(2) Data and methods

2-2) Methods

2-2-1) A brief review

- Main reference – Verbelen R. et al. (2022)
- Borrow wisdom from actuarial science
- Chain Ladder (CL) an industrial standard
- Mack's non-parametric approach & Poisson and negative binomial

2-2-2) Chain Ladder setup

- We need start date s and evaluation date τ, to delineate our observational window
- N_t: total number of the groups that arrived in the t^{th} week
- N^r_t: total number of the groups that arrived in the t^{th} week and registered
- $N^r_t = N^r_{td}$ total number of the groups that arrived in the t^{th} week and registered within the observational window ($d < \tau$)

(2) Data and methods

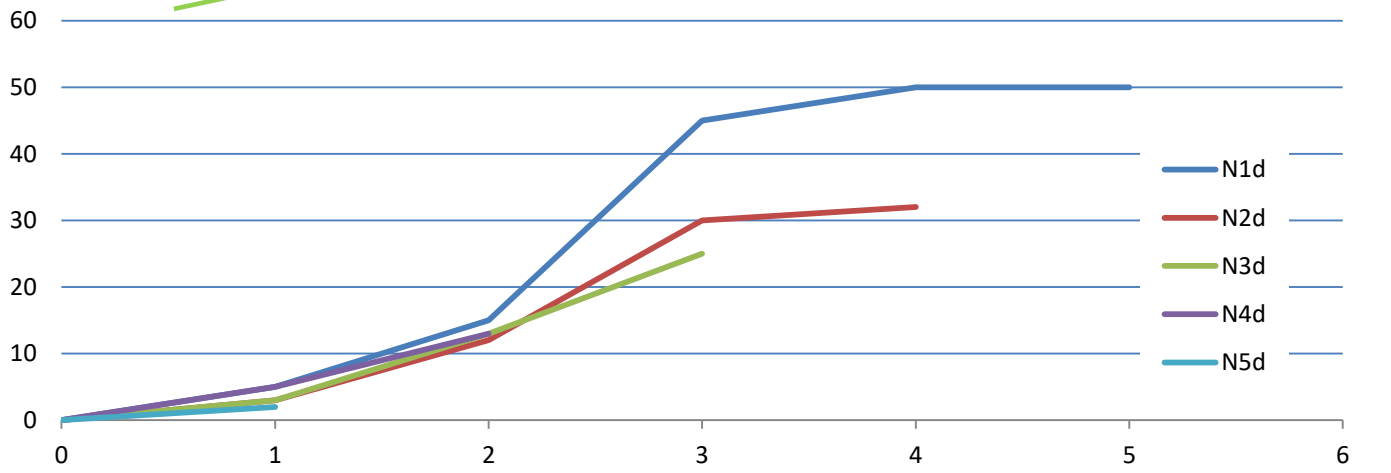
2-2) Methods

2-2-2) Chain Ladder setup => Mack's CL

Cumulative Triangle

	Registration Weeks				
Arrival Week	**1**	**2**	**3**	**4**	**5**
"2022-03-05"	5	15	45	50	50
"2022-03-12"	3	12	30	32	
"2022-03-17"	3	13	25		
"2022-03-24"	5	13			
"2022-03-31"	2				

N_1
N_2
N_3
N_4
N_5



(2) Data and methods

2-2) Methods

2-2-3) Chain Ladder setup => Poisson family

Incremental Triangle

	Registration Weeks				
Arrival Week	**1**	**2**	**3**	**4**	**5**
"2022-03-05"	5	3	3	5	2
"2022-03-12"	10	9	10	8	
"2022-03-17"	30	18	12		
"2022-03-24"	5	2			
"2022-03-31"	0				

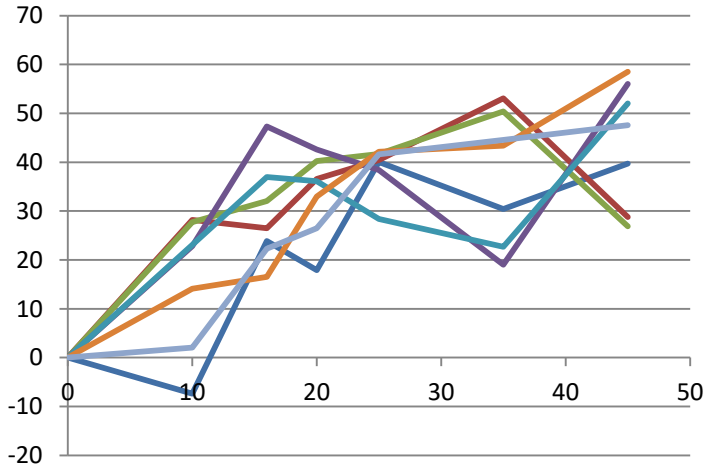
Arrival Week	Counts	Dev.Week	Other Covariates…
"2022-03-05"	5	1	
"2022-03-05"	10	2	External
"2022-03-05"	30	3	Factors
"2022-03-05"	5	4	**Time Effect**
"2022-03-05"	0	5	
"2022-03-12"	3	1	
"2022-03-12"	9	2	or to group our data by regions
"2022-03-12"	18	3	to use location Information
"2022-03-12"	2	4	for future analyses
…	…	…	…

(3) Results

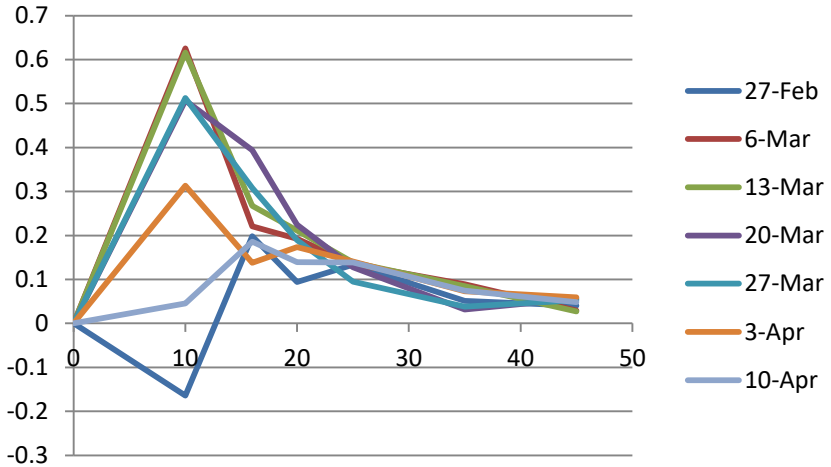
3-1) Profile of error counts and error percentage

Selection of the start week and τ to determine the observational window

Example of applying Mack's CL to evaluate the ultimate counts error rate



Total error percentage

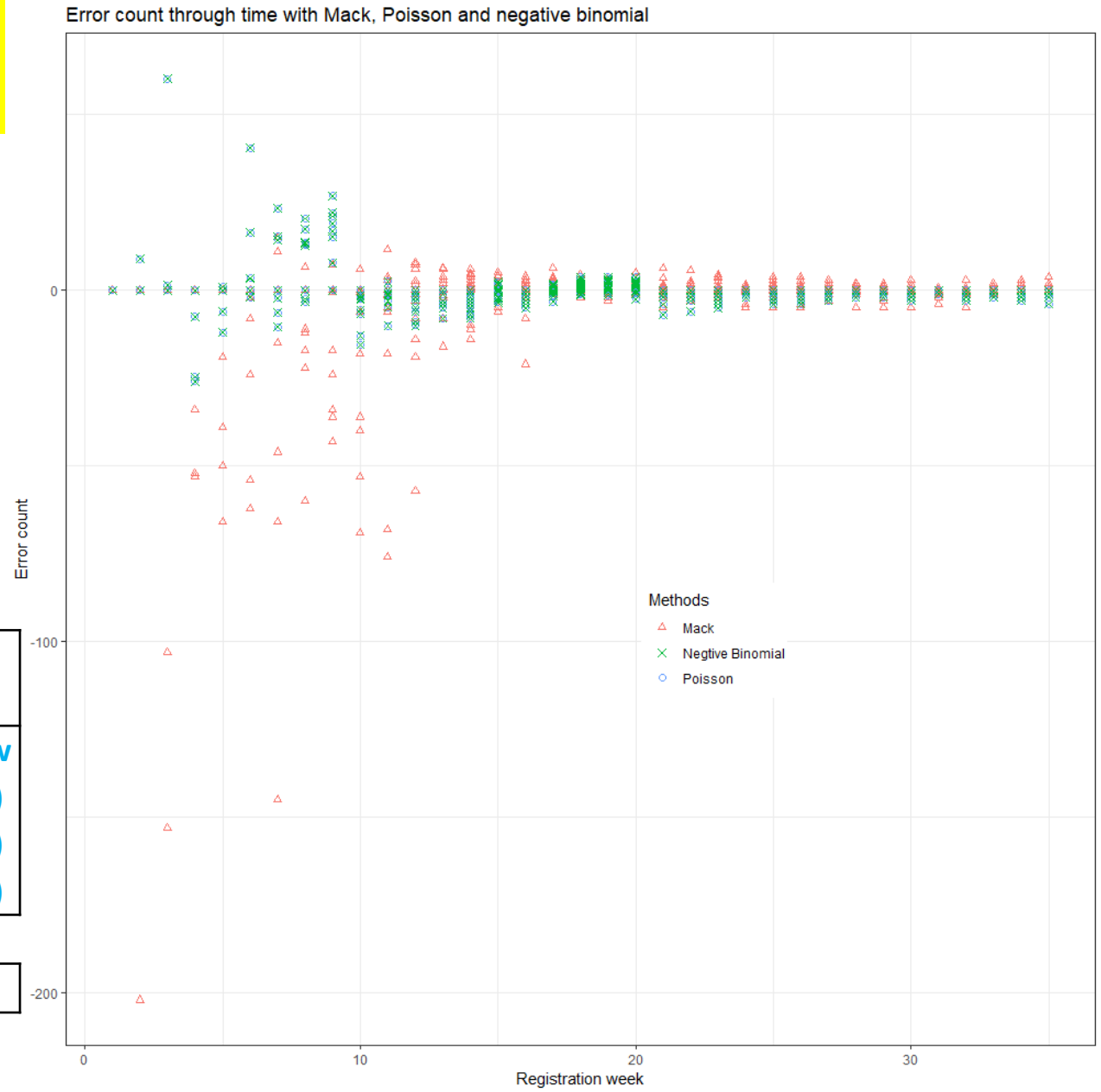


Error percentage per cell in the lower triangle

Observational windows: 10, 16, 20, 25, 35, 45 weeks

(3) Results

3-2)
The negative binomial model works the best



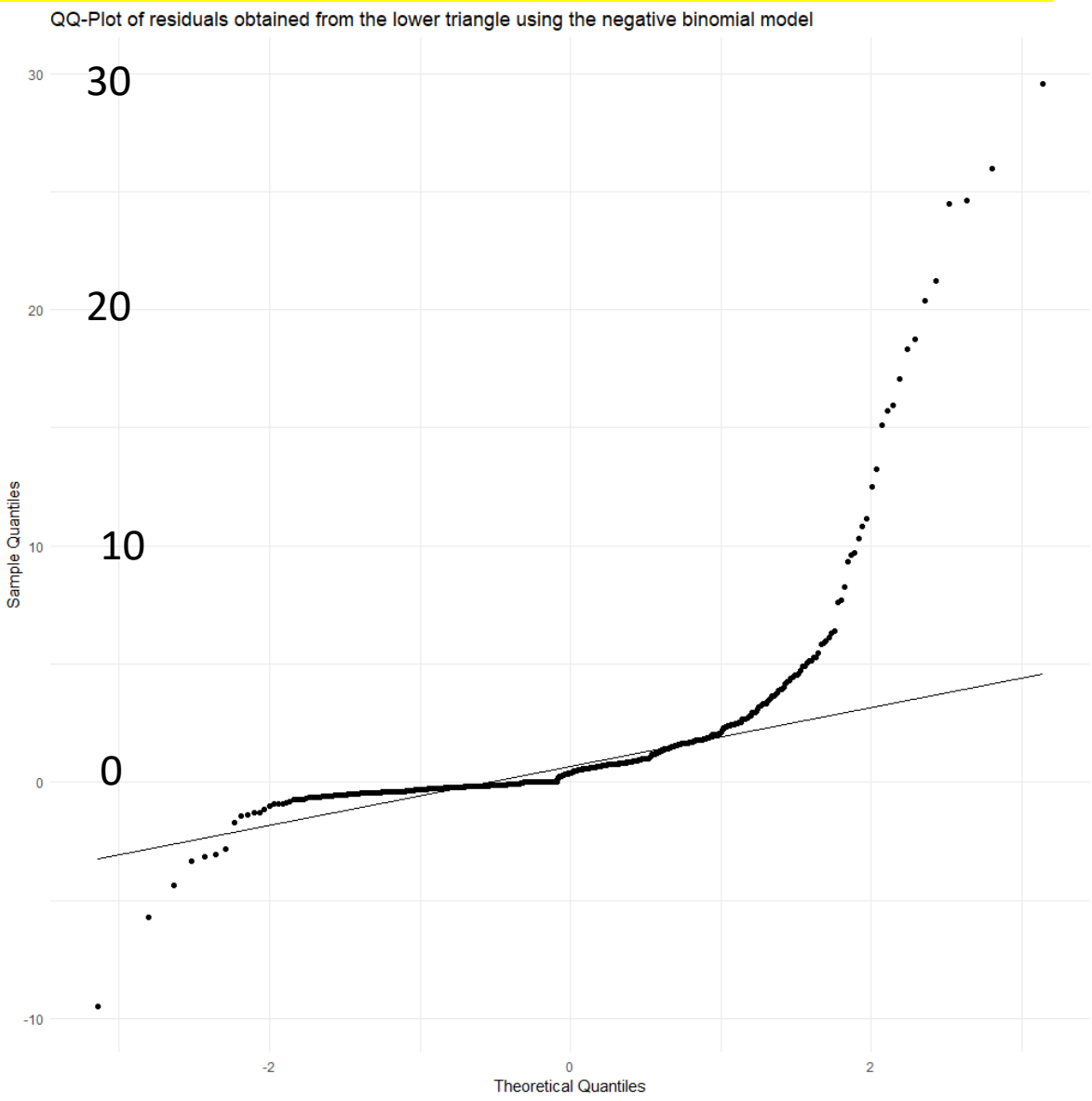
Start = 2022-Mar-06 Window length = 35 weeks		
Error	**Ultimate**	**In Window**
Mack	917 (53%)	773 (49%)
Poisson	917 (53%)	773 (49%)
Neg. Bi.	882 (51%)	738 (47%)

Unobserved	1858	1729

(3) Results

3-2)

The negative binomial model works the best. But

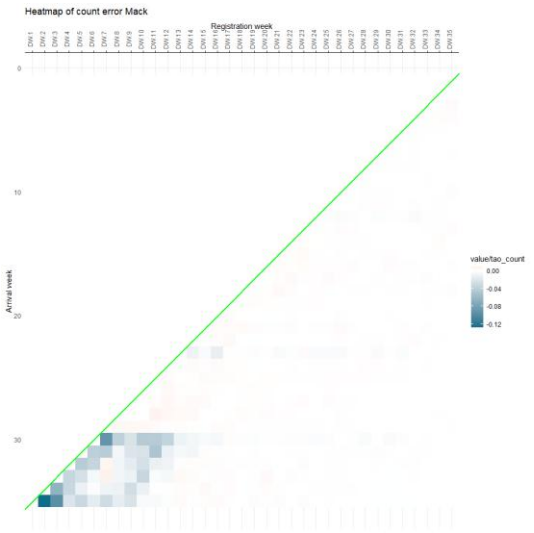


Start = 2022-Mar-06 Window length = 35 weeks		
Error	**Ultimate**	**In Window**
Mack	917 (53%)	773 (49%)
Poisson	917 (53%)	773 (49%)
Neg. Bi.	882 (51%)	738 (47%)

Unobserved	1858	1729

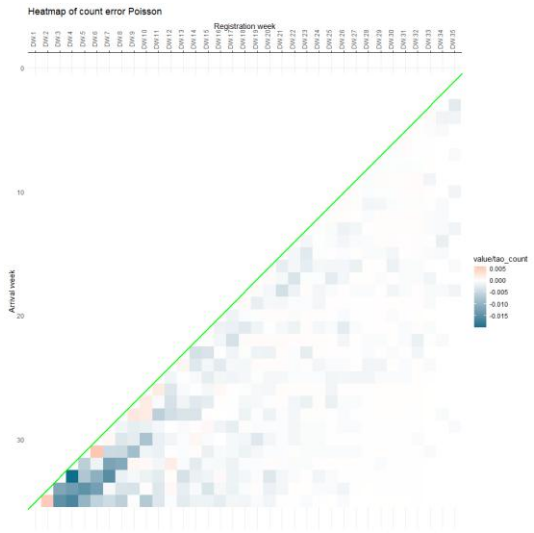
(3) Results

3-3) For administrative or policy-making purposes, maybe knowing the patterns of errors distributed is more important



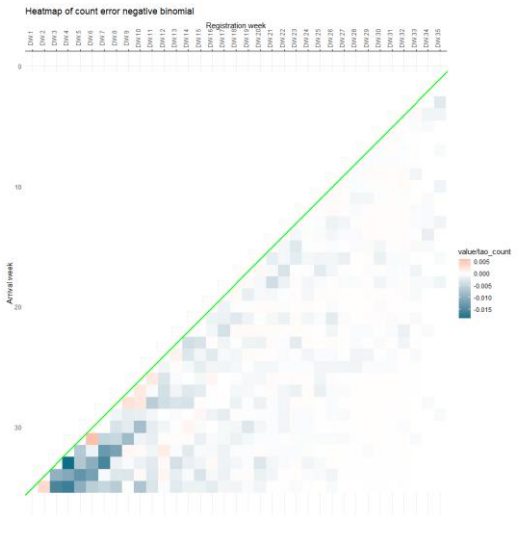
Mack

773 (49%)



Poisson

773 (49%)



Negative Binomial

738 (47%)

(3) Results

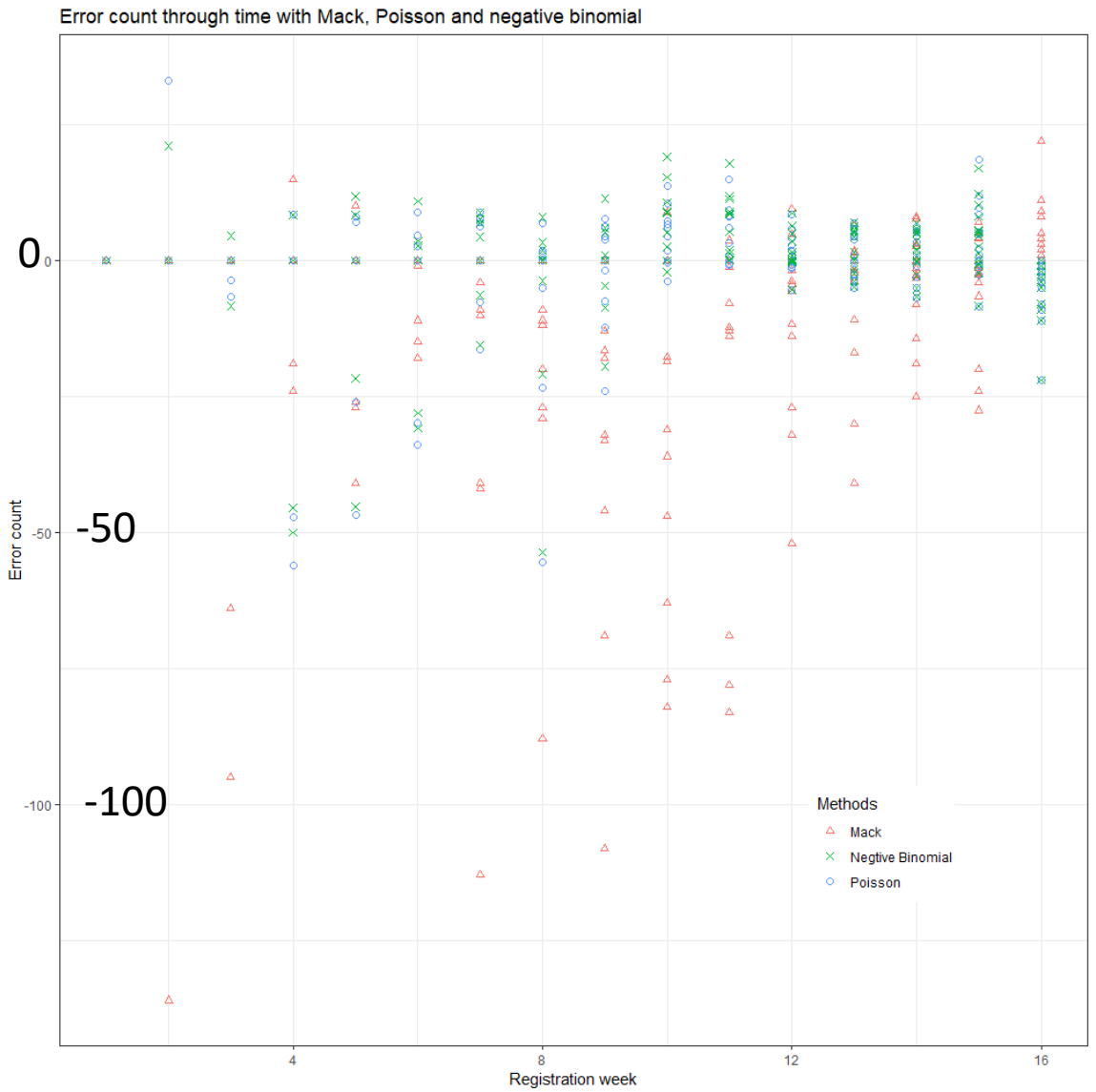
3-3) The ultimate errors are high, but the weekly errors are low, if we applied generalized linear models

Future week	Mack		Poisson		Negative binomial	
	Counts error	Error rate	Counts error	Error rate	Counts error	Error rate
1	-202	-12.74	9	0.54	6	0.36
2	-256	-16.15	-42	-2.67	-40	-2.50
3	-139	-8.77	-78	-4.95	-73	-4.58
4	-174	-10.98	-60	-3.79	-58	-3.66
5	-150	-9.46	-39	-2.49	-37	-2.33
6	-246	-15.52	-59	-3.69	-62	-3.90
7	-127	-8.04	-51	-3.20	-50	-3.14
8	-183	-11.55	-23	-1.42	-21	-1.30
9	-217	-13.67	-23	-1.47	-23	-1.45
10	-157	-9.90	-30	-1.88	-30	-1.91
11	-76	-4.81	-17	-1.08	-16	-1.01
12	-1	-0.04	-23	-1.44	-21	-1.31

(3) Results

3-4) Taking 16 weeks from 2022-Mar-06

Within window error percentage		
Mack	Poisson	Neg. Bin.
-6.83	1.66	1.06
-7.99	-0.51	-0.19
-1.41	-4.76	-4.38
-4.22	-2.89	-2.35
-3.02	-2.39	-2.10
-11.01	0.31	0.28
-9.85	-3.68	-3.27
-16.86	-1.18	0.11
-18.26	2.31	3.88
-14.46	2.78	3.80
-6.95	0.79	1.03
-5.07	1.05	1.11
-2.78	0.76	0.90
-3.96	2.96	3.00
3.47	-3.47	-3.47



(4) Conclusion

(1) We can apply generalized linear models to nowcast the weekly arriving groups of immigrants.

(2) The methodology provided by Verbelen R. et al. (2022) can be extended to automation implementation.

(3) The factors that affect administrative delays should be further studied.

Statistics Flanders Research Seminar

Thank you for listening

Reference:

Roel Verbelen, Antonio, K., Claeskens, G. & Crevecoeur, J. Modeling the Occurrence of Events Subject to a Reporting Delay via an EM Algorithm. Statistical Science 37, (2022).

Lie HONG

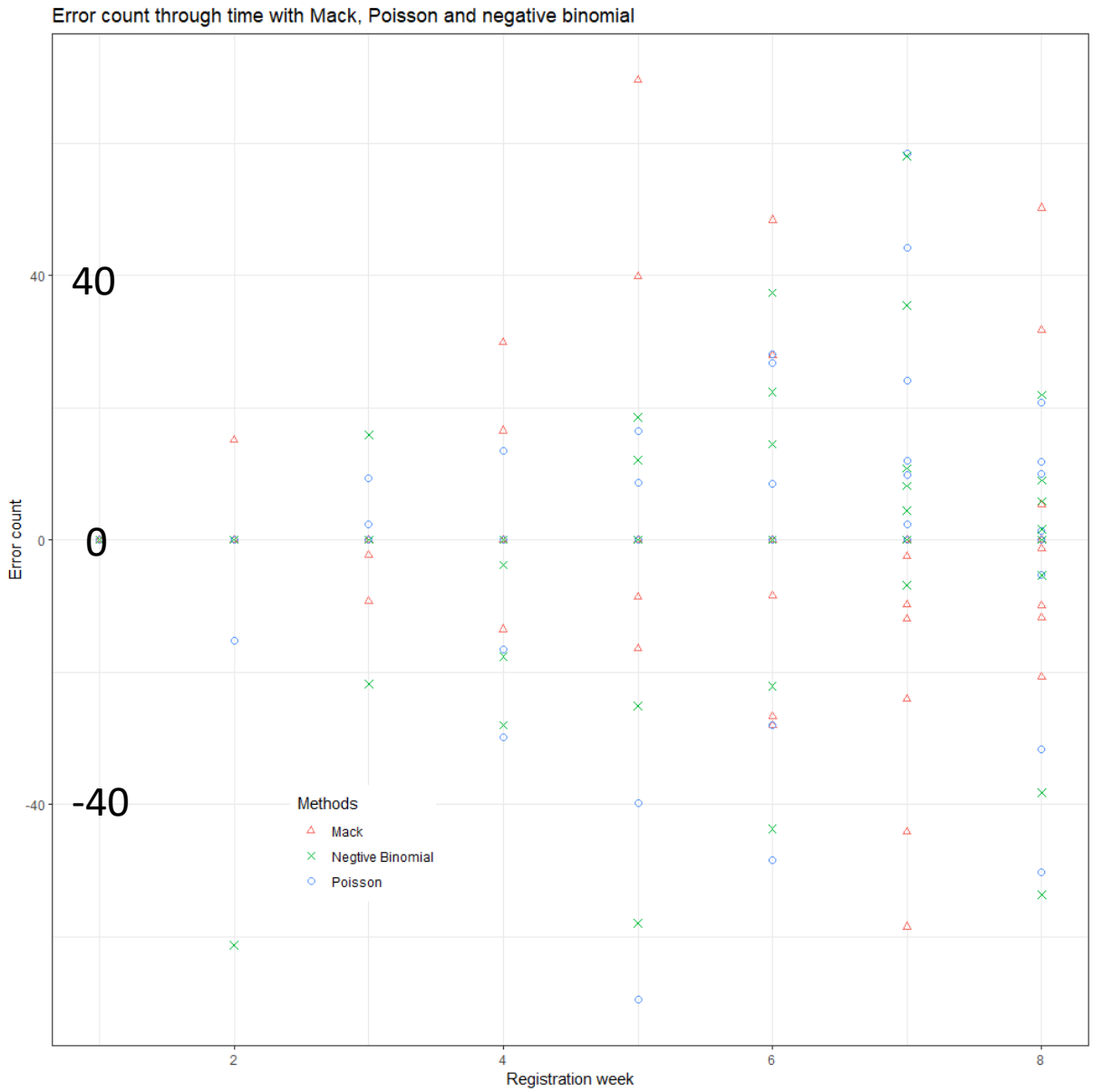
Supervised by Prof. Dr. Katrien Antonio & Dr. Jorre Vannieuwenhuyze

2024-May-21

(3) Results

3-4) Taking 8 weeks from 2022-Feb-27

Negative binomial	
Counts error	Error rate
0	0.00
-15	-0.39
12	0.30
-33	-0.85
-84	-2.19
-13	-0.34
151	3.92
-43	-1.13



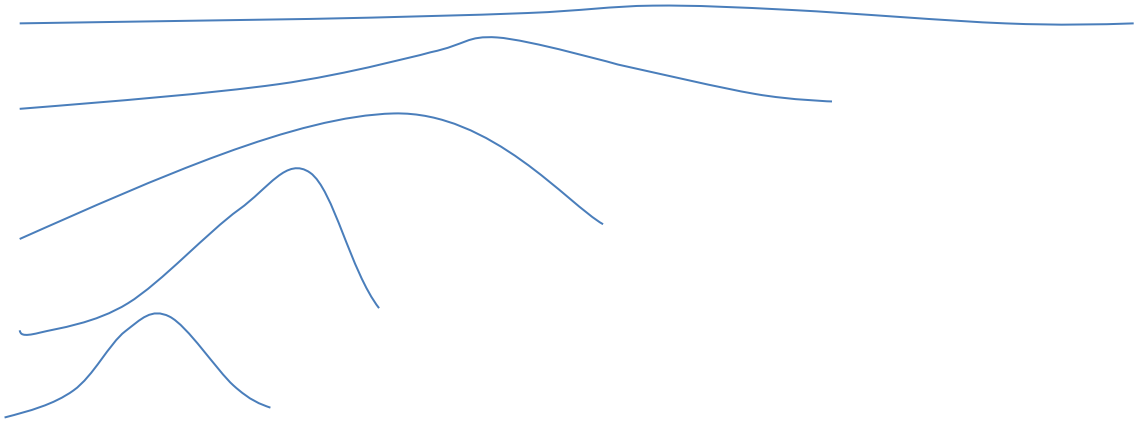
(2) Data and methods

2-2) Methods

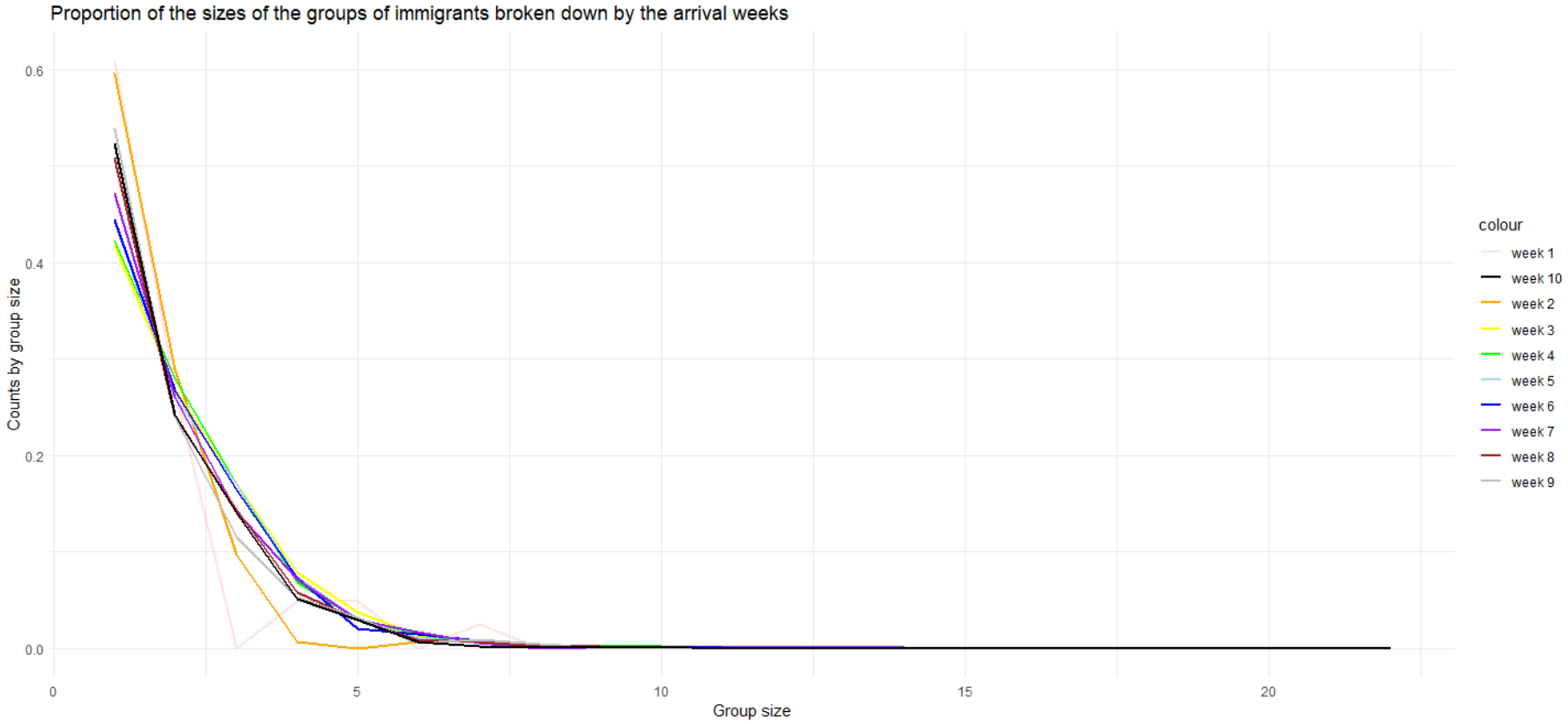
2-2-2) Chain Ladder setup => Poisson family

Incremental Triangle

	Registration Weeks				
Arrival Week	**1**	**2**	**3**	**4**	**5**
"2022-03-05"	5	3	3	5	2
"2022-03-12"	10	9	10	8	
"2022-03-17"	30	18	12		
"2022-03-24"	5	2			
"2022-03-31"	0				



(2) Data and methods



(2) Data and methods