Voorstel voor een methode voor een stochastisch regionaal prognose

Trond

08 July, 2020

1 Aanleiding

De huidige methodiek gebruikt om bandbreedtes in PEARL te bepalen is quasi-stochastisch. Er zijn al eerder pogingen geweest om een volle stochastische methodiek op te stellen, deze pogingen zijn echter gestrand wegens gebrek aan tijd en computercapaciteit. Door de ervaring van de proefrun met Microsoft Azure vorig jaar hebben we geleerd dat het draaien van PEARL in de cloud goed te doen is. Daarom lijkt het nuttig om opnieuw te kijken naar de methodiek voor een beperkte stochastische run.

Een toegevoegde waarde van een volle stochastische methode tov de huidige quasistochastische methode is dat alle voorspellingsvariabelen, inclusief ratio's en andere afgeleide indices, probabilistisch consistent zijn. Voorspellingen kunnen worden gemaakt door middel van relatief eenvoudige Monte Carlo-simulaties om een verdeling van mogelijke resultaten te produceren. Het is dan noodzakelijk om voor elke demografische component de gemiddelde (of mediaan), variantie-covariantiestructuur en distributievorm te specificeren. Deze parameters kunnen worden geschat op basis van "expert judgment", door evaluatie van schattingsfouten of door een tijdsreeksmodel (Keilman 2001).

Hier ontwikkelen we een methode voor probabilistische voorspellingen van vertrek en woningbouwproductie op basis van tijsdreeksmodellen. De aanpak is gebaseerd op het Lee Carter model, vaak gebruikt om leeftijdspecifieke sterfte of sterftekans te schatten (Booth et al. 2006). Het Lee Carter model is inmiddels ook voor migratie toegepast (Wiśniowski et al. 2015), voor zowel verhuiskansen als voor aantallen. Voor aantallen is de Poisson versie van het Lee Carter model het meest geschikte (Brouhns, Denuit, and Vermunt 2002). Lee Carter modellen worden meestal gebruikt voor nationale prognoses, en voor demografische variabelen met een leeftijdsdimensie. Hier gebruiken we ze ook voor variabelen met een regio-dimensie, wat, zover ik weet, niet eerder is gedaan.

In deze notitie laat ik zien hoe we Lee Carter modellen kunnen gebruiken om de volgende PEARL-input te voorspellen: regio-specifieke aanpassingsfactoren zoals gebruikt in PEARL en regio-specifieke woningvoorraad waarvan we de netto woningbouwproductie kunnen berekenen. Als proof of concept worden de voorgespelde variabelen gecombineerd met een versimpelde cohort-component model voor een Monte Carlo voorspelling van de toekomstige netto migratie en risikobevolking.

2 Binnenlandse migratie in PEARL

In PEARL wordt het aantal vertrokken mensen uit een gemeente in meerdere stappen bepaald. Hieronder een korte omschrijving van deze stappen, gemaakt op basis van Andries' paper over PEARL.

Het aantal nationale verhuizingen en risikobevolking voor leeftijd x en geslacht g in jaar t, schrijven we als $M^g_{x,t}$ en $P^g_{x,t}$. De nationale leeftijdspecifieke verhuiskans is berekent op de volgende manier

$$V^g_{x,t} = M^g_{x,t}/P^g_{x,t} \tag{1}$$

Het aantal vertrokken personen uit gemeente i, $\hat{M}^g_{i,x,t}$, wordt in eerste instantie geschat als het product van de nationale verhuiskans en de risikobevolking in gemeente i, $P^g_{i,x,t}$.

$$\hat{M}^g_{i,x,t} = V^g_{x,t} * P^g_{i,x,t} \tag{2}$$

Dit wordt dan vervolgens gecorrigeerd met een regionaal factor $F^g_{i,t}$. Deze factor is berekent als

$$F^g_{i,t} = \frac{\bar{M}^g_{i,t}}{\sum_x \hat{M}^g_{i,x,t}} \tag{3}$$

waar $\bar{M}^g_{i,t} = (1/3)\sum_{z=1}^{3} M^g_{i,x,t-z}$, ofwel het gemiddelde aantal vertrokken personen de afgelopen drie jaar. Het uiteindelijke aantal vertrokken personen wordt bepaald als

$$M^g_{i,x,t} = F^g_{i,t} * \hat{M}^g_{i,x,t} \tag{4}$$

In PEARL is er een nauwe samenhang tussen netto nieuwbouw en vestiging in een regio, maar er zijn in het model meerdere manieren waarop nieuwbouw vestiging beinvloedt. In deze notitie maken we geen onderscheid tussen lange- en korte-afstandsverhuizingen, maar gaan er simpelweg vanuit dat een er positieve relatie bestaat tussen nieuwbouwwoningen en vestigingskans.

3 Stochastisch opzet PEARL

Ik stel voor om de regionale factor, $F^g_{i,t}$ en het woningvoorraad $W_{i,t}$ te varieren. Op deze manier kunnen we twee verschillende effecten simuleren: variantie aan de vestigingskant via $W_{i,t}$ en variantie aan de vertrekkant via $F^g_{i,t}$.

3.1 Lee Carter modellen

In de literatuur worden Lee Carter modellen gebruikt voor leeftijd-specifieke processen, maar hier gebruiken we ze ook voor processen met een regio-dimensie. Dit betekent dat d bevat zowel leeftijd x als regio i. De processen die we met het model beschrijven zijn de regionale factor en woningbouwproductie per regio. Het Lee-Carter model wordt verder gespecificeerd als een functie van de regio-specifieke parameters α_i (level) en β_i (slope) en de tijdspecifieke parameters k_t.

De regionale factor is, in alle regio's bijna gelijk voor mannen en vrouwen. Daarom schatten we èèn Lee Carter model voor de regionale factor, met verschillend intercept δ_g voor mannen en vrouwen. Het model gebruikt voor de schatting van de regionale factor is een generalised linear Gaussian model log link. Het model kunnen we als volgt schrijven:

$$\begin{aligned} \log \lambda_{i,g,t} &= \alpha^F_i + \delta^F_g + \beta^F_i k^F_t + \varepsilon^F_{i,g,t} \\ F_{i,g,t} &\sim \mathcal{N}\left(\lambda_{i,g,t},\, \varepsilon^F_{i,g,t}\right) \end{aligned}$$

Het Lee Carter model voor woningbouw wordt geschat als een generalised linear Poisson model met log link. Het model kunnen we als volgt schrijven:

$$\begin{aligned} \log \mu_{i,t} &= \alpha^W_i + \beta^W_i k^W_t + \varepsilon^W_{i,t} \\ W_{i,t} &\sim \text{Poisson}\left(\mu_{i,t}\right) \end{aligned}$$

Hoe zit dit allemaal in elkaar? Een klein voorbeeld met een een-stap-vooruit voorspelling is wellicht handig. Eerst worden de Lee Carter modellen geschat, waardoor we schattingen van de parameters en de pure tijdstrend krijgen. Vervolgens maken we een een-stap-vooruit voorspelling van $\hat{k}_{t+1}$ en we kunnen de regionale factor en nieuwbouwwoningen in $t+1$ berekenen met de geschatte parameters van de Lee Carter modellen als

$$\hat{\lambda}_{i,g,t+1} = \exp\left(\hat{\alpha}_i^F + \hat{\delta}_g^F + \hat{\beta}_i^F \hat{k}_{t+1}^F\right)$$
$$\hat{\mu}_{i,t} = \exp\left(\hat{\alpha}_i^W + \hat{\beta}_i^W \hat{k}_{t+1}^W\right)$$

3.2 Schattingsmethode

We schatten de Lee Carter modellen als non-lineaire systemen, gebruikmaakend van het gnm-pakket in R (Turner and Firth 2007). Zie ook de discussie in Currie (2013). De methodes hier gekozen hebben hun beperkingen maar zijn vrij recht-aan-recht toe en kant-en-klare software is beschikbaar.

Omdat de eenheden van de afhankelijke variabelen wat verschillen zijn de Lee Carter modellen gefit onder net wat verschillende aannames: de verhuizing- en woningbouwmodellen worden geschat als Poisson met log link en risiko-bevolking als offset variabel, en de regionale factor modellen worden geschat als Gaussian met log link.

3.3 Voorspellingen

Als we zien van de vergelijkingen boven wordt de dynamiek in de tijsreeks volledig opgevangen door de geschatte pure tijdstrend k_t. k_t is een vector met de volgende 2 elementen

$$k_t = \begin{bmatrix} k_t^F \\ k_t^W \end{bmatrix}$$

Om k_t te kunnen voorspellen maken we gebruik van een univariate exponential smoothing model (Hyndman et al. 2002). Hier gebruiken we de functie *ets()* in R waarmee het volgende single-source of error state space model wordt geschat

$$k_t = \theta_{t-1} + e_t$$
$$\theta_t = \theta_{t-1} + \nu e_t$$

De schattingsfouten e_t zijn normaalverdeeld met variantie σ^2:

$$e_t \sim NID(0, \sigma^2)$$

De twee geschatte parameters hier zijn ν en σ^2. Als $\nu = 1$ is dit een random walk. De vergelijking hierboven beschrijvt dus een simpel stochastisch systemen, waarmee we tokoemstige waarden van k_t^g kunnen simuleren. Een h-stap-vooruit voorspellingen van simulatie s kunnen we schrijven als $\bar{k}_{t+h,s}$. Op basis van de simulaties s kunnen we gemiddelde waarden, $\hat{k}_{t+h}$, en betrouwbaarheidsintervallen berekenen.

4 Proefrun met een dummy model

In deze sectie worden resultaten van een proefrun met het model beschreven. Hoofddoel van de exercitie is te analyseren hoe gevoelig de voorgespelde risikobevolking is ten opzichte van onzekerheden in (i) vertrek en (ii) vertrek en woningbouwproductie gecombineerd. De opzet is vergelijkbaar met Hyndman and Booth (2008),

die een Monte Carlo-simulatie van toekomstige bevolking naar geslacht uitvoeren, door Lee Carter modellen voor vruchtbaarheid, sterfte en netto migratie te combineren met en een cohort componentmethode.

In deze notitie maken we een Monte Carlo analyse van de risikobevolking met 100 simulaties van k_t tot 2040. Het genereren van simulaties is vrijwel volledig automatisch en het model schaalt daarom heel makkelijk; op een computer met meer geheugen kunnen we meer dan 100 simulaties draaien zonder dat dat ekstra werk oplevert.

De data gebruikt om de Lee Carter modellen te schatten betreeft tijdsreeksen over verhuizingen, bevolking en woningbouw tussen 1995 en 2017 op gemeenteniveau gebruikt; alles beschikbaar van CBS Open Data. Het gaat over de volgende series: migratie 60048ned, bevolking 70072ned, woningbouw 37713. We gebruiken de leeftijdsgroepen uit de series 60048ned en gemeente als regionaal niveau. Alle reeksen zijn gecorrigeerd voor gemeentelijke herindelingen volgens de methodiek gebruikt door het PBL en het CBS. Verhuizingen worden gedefinieerd als de som van binnenregionale verhuizingen en vertrek - net als in PEARL.

Voor de tijdsreeksen van $\bar{M}^g_{i,t}$ en $W_{i,t}$ worden 3-jaar voortschrijdend gemiddelde gebruikt, wat betekent dat de tijdsreeks voor de regional factor en voor nieuwbouwwoningen van 1998 tot 2017 loopt. De risikobevolking is berekent als het gemiddelde van de bevolking aan het begin en aan het einde van het jaar.

De kant van vestiging wordt zo simpel mogelijk gehouden; in een volschala run zal vestiging immers worden bepaald met PEARL. In deze notitie bereken ik de vestigingskant $\pi_{i,t}$ in regio i als het aandeel woningen van de gemeente in het totale woningvoorraad

$$\pi_{i,t} = \frac{W_{i,t}}{\sum_j W_{j,t}}$$

Dan kunnen we het model van de vergelijkingen (1) - (4) sluiten met de volgende regel voor dynamische ontwikkeling van de risikobevolking:

$$P^g_{i,x,t+1} = P^g_{i,x,t} - M^g_{i,x,t} + \sum_j M^g_{j,x,t} * \pi_{i,t} \tag{5}$$

4.1 Simulaties van de tijdstrend

Van de omschrijving boven weten we dat de vector k_t 2 geschatte tijstrenden bevat: regionale factor en woningbouwproductie. De resultaten van de simulaties zijn te zien in Figuur 1. De zwarte lijn in de figuren zijn de $\hat{k}_t$ geschat met de Lee Carter modellen en de grijze lijnen zijn h-stappen-vooruit simulaties gemaakt met het tijdsreeks model, $\hat{k}_{t+h}$.

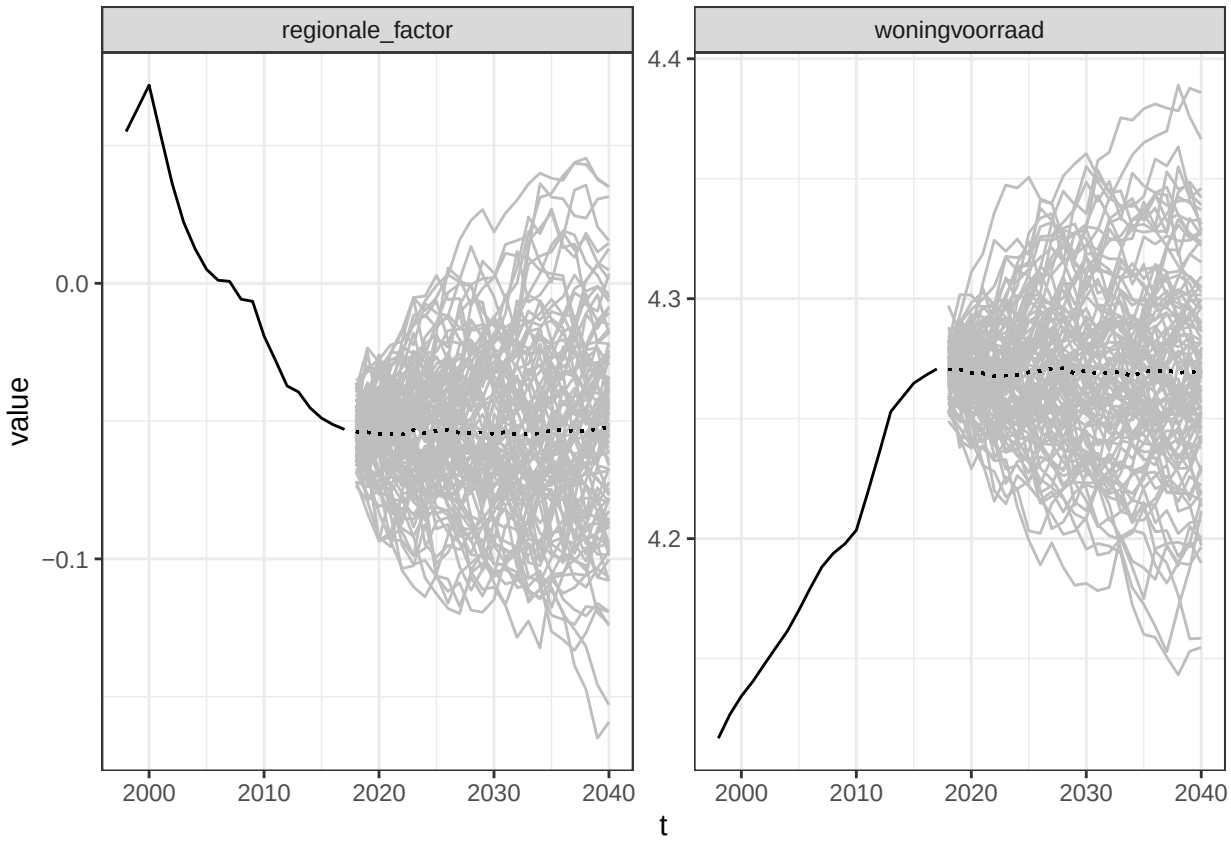

Figure 1: Simulaties van de tijdstrend

4.2 Resultaten van het model van vergelijkingen (1) tot (5)

De simulaties van de vorige sectie worden gecombineerd met het migratiemodel in vergelijkingen (1) - (5) om de ontwikkeling van de risikobevolking tot 2040 te simuleren. Om de resultaten te laten zien kiezen we als voorbeeld de vier gemeenten van de groote steden. De grijze vlakken in de figuren geven een 80 % bandbreedte aan en de vaste lijn het gemiddelde van alle simulaties.

Om de relatieve invloed van de verschillende variabelen (regionale factor; nieuwbouwwoningen) op de voorgespelde risikobevolking te ontrafelen kunnen we switches introduceren. Deze swiches verzamelen we in een vector $\gamma = \{\gamma_f, \gamma_w\}$ van binaire variabelen. De h-stap-vooruit voorspelling, $\tilde{k}_{t+h,s}$ kunnen we definieren als

$$\tilde{k}_{t+h,s} = \gamma \hat{k}_{t+h,s} + (1 - \gamma)\bar{k}_{t+h}$$

Van deze vergelijking zien we dat $\gamma_f = 0$ bijvoorbeeld betekent dat er met de middenvariant van de regionale factor wordt gewerkt en $\gamma_f = 1$ dat we de hele waaier van gesimuleerde waarden van Figuur 1 gebruiken.

Figuur 2 laat de netto migratie (som over leeftijden) in de vier grote steden zien. In onze nep-model hebben de vier grote steden een negatieve saldo. Links in het figuur waarin $\gamma_f = \gamma_w = 0$ wordt voor alle drie variabelen met een middenvariant gewerkt. Niet geheel verrassend is er dan ook geen variatie in de uitkomsten in dit scenario. Het figuur laat zien dat variatie in de regionale factor over het algemeen het grootst effect heeft op de netto migratie, en dat de effecten verschillen tussen de steden. Aan de rechterkant van het figuur zien ook

een bepaalde stapeling van onzekerheid, die ook nog enigszins regionaal verschilt. Het valt op dat het model zich beweegt richting een evenwicht zonder netto migratie - dit is natuurlijk anders in PEARL.

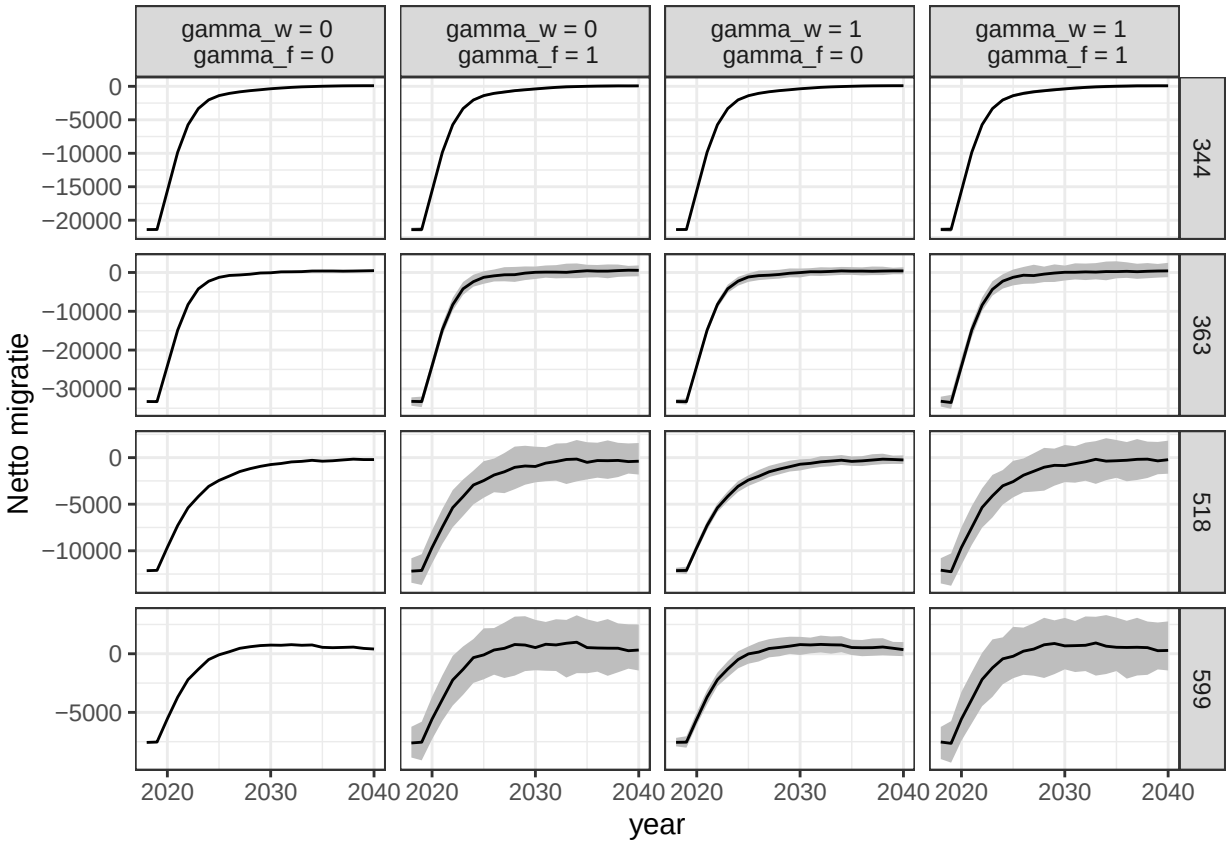

Figure 2: Netto migratie (totaal) in s' Gravenhage (518), Amsterdam (363), Utrecht (344) en Rotterdam (599)

De ontwikkeling van de (totale) risikobevolking in de vier groote gemeenten is te zien in Figuur 3. De zwarte lijnen in het figuur is de gemiddelde voorspelling en het grijz gebied om de lijn 80 procent onzekerheidsmarges. Aan de linkerkant in de figuur heeft woningbouw geen invloed op vestiging ($gamma_w = 0$) en we zien grote verschillen tussen de steden. Ondanks dat de regionale factor groter effect heeft op de risikobevolking dan woningbouw, worden de bandbreedtes wel groter als we woningbouw ook laten varieren.

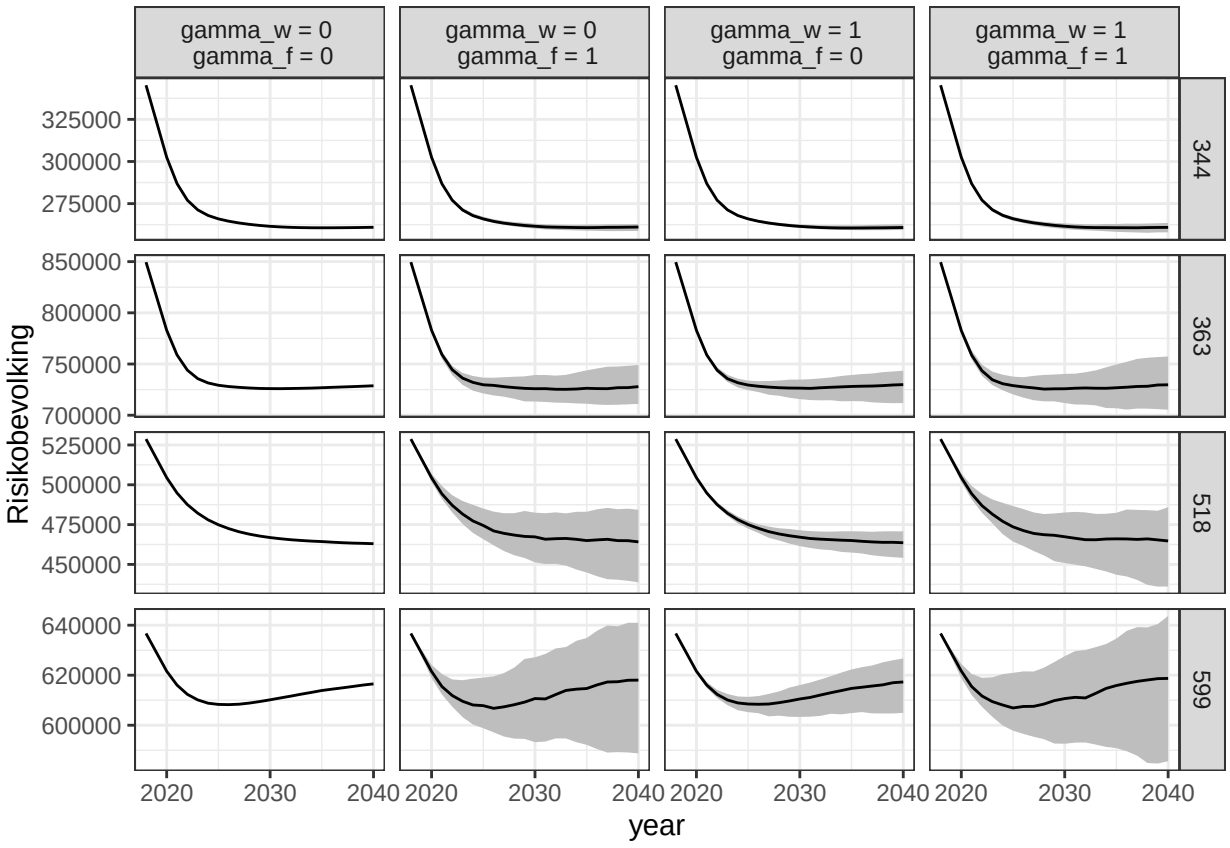

Figure 3: Risikobevolking (totaal) in s' Gravenhage (518), Amsterdam (363), Utrecht (344) en Rotterdam (599)

Figuur 4 laat resultaten per leeftijdsgroep, voor jaren 2020, 2030 en 2040 zien. We zien dezelfde dynamiek terug als in de voorafgaande figuur: te groter invloed van woningbouw op de vestigingskans, te groter de onzekerheidsmarges. We zien ook wat verschuivingen in de leeftijdsverdeling. Ik ga hier niet verder in op de oorzaken daarvan, maar dit is juist interessant om te onderzoeken in een meer diepgaande analyse.

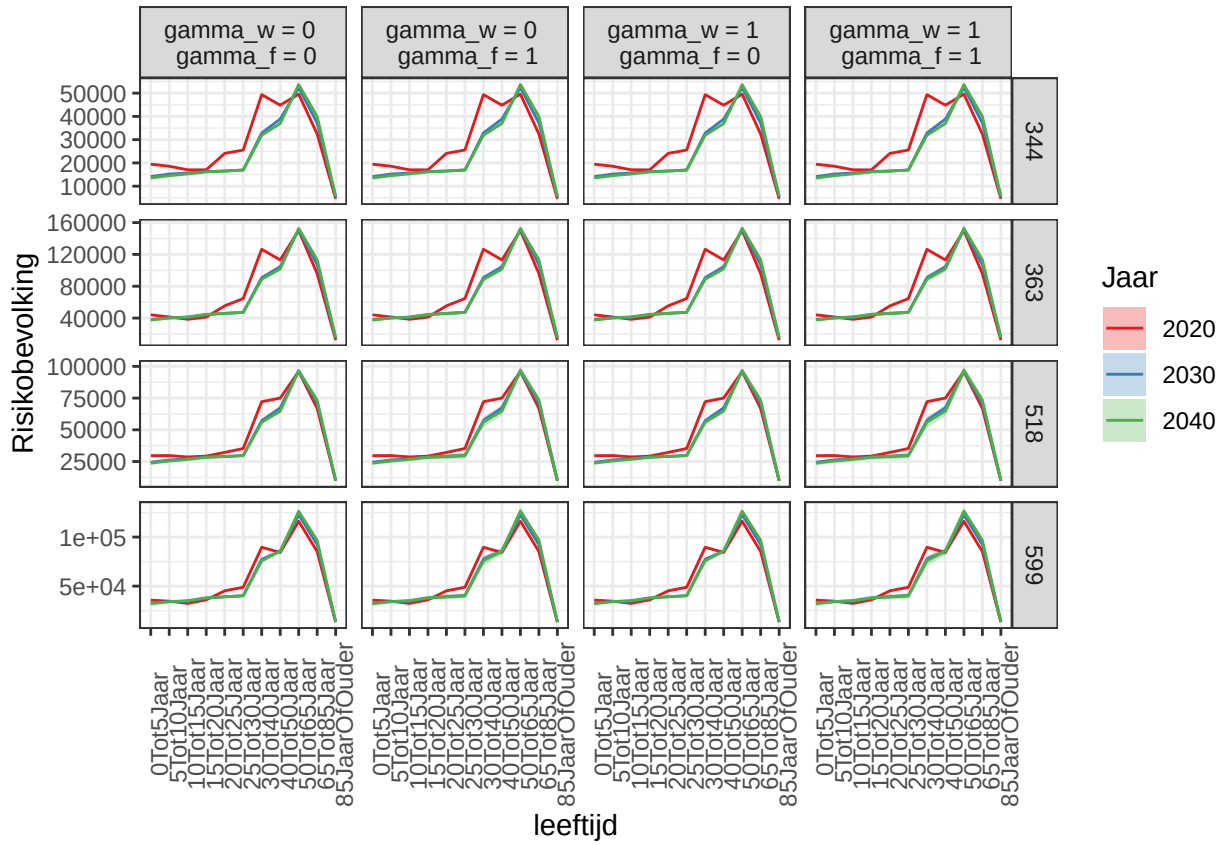

Figure 4: Leeftijdsverdeling van de risikobevolking in s' Gravenhage (518), Amsterdam (363), Utrecht (344) en Rotterdam (599)in 2020, 2030 en 2040

5 Conclusies, opmerkingen en hoe verder

In dit document heb ik een voorstel gemaakt voor en stochastische simulatiemethode geschikt voor PEARL. De groote voordeel van deze methode is de dimensionaliteitsreductie: de test die ik hier heb gegraaid kunnen we bijna direct toepassen in PEARL. Zover ik weet is het Lee Carter model nooit toegepast op regionale prognoses, in die opzicht is dit ook nieuw in de literatuur.

De resultaten hier suggereren dat de meeste variatie in netto migratie en risikobevolking kwam door variatie verandering in de regionale factor, maar ook dat de onzekerheid toenam als we zowel de regionale factor als nieuwbouw tegelijk varieerden. Deze resultaten zijn uiteraard een gevolg van de eenvoudige formule voor update van de bevolking. In PEARL, waar woningbouwproducie een ander rol heeft, is de relatieve invloed van die twee factoren anders.

Dit document moet worden gezien als een 'proof-of-concept'. De resultaten in de vorige sectie laten met name zien hoe het zou kunnen werken in de context van PEARL, maar door de aannames zijn de resultaten ook niet noodzakelijk realistisch. De geschatte Lee Carter modellen zouden we beter naar moeten kijken. Er zijn een aantal andere schattingsmethode's die wellicht nauwkeuriger zijn dan die die ik hier heb gebruikt. Bijvoorbeeld Structural Value Decomposition is een andere bekende optie hiervoor (Stoeldraijer et al. 2013).

Een uitdaging met de Monte Carlo method zoals hier voorgesteld is dat de bandbreedtes mogelijk erg groot kunnen worden (Renshaw and Haberman 2008). Een optie is om een meer gerichte Monte Carlo, waarin stochastische simulaties worden gecombineerd met expert judgment. Hierbij een suggestie van de

PBL-thuisexpert Andries: “Trek voor een bepaalde regio een bepaalde waarde uit de verdeling van de regionale factor (en hou voor de andere regio’s de regionale factor gelijk aan de prognose) en trek voor de omliggende gemeenten (van die specifieke regio) uit de verdeling van de woningbouw. Deze waarden kun je dan in PEARL stoppen. Per regio zou je dit volgens de principes van Monte Carlo modelering een flink aantal keren moeten doen, bv. 50 keer. Dit betekent dat je dan - in de cloud - PEARL het aantal regio’s (in principe alle gemeenten) * 50 keer moeten runnen (uitgaande van b.v. 400 gemeenten dus rond 20 duizend runs).”

6 Appendix

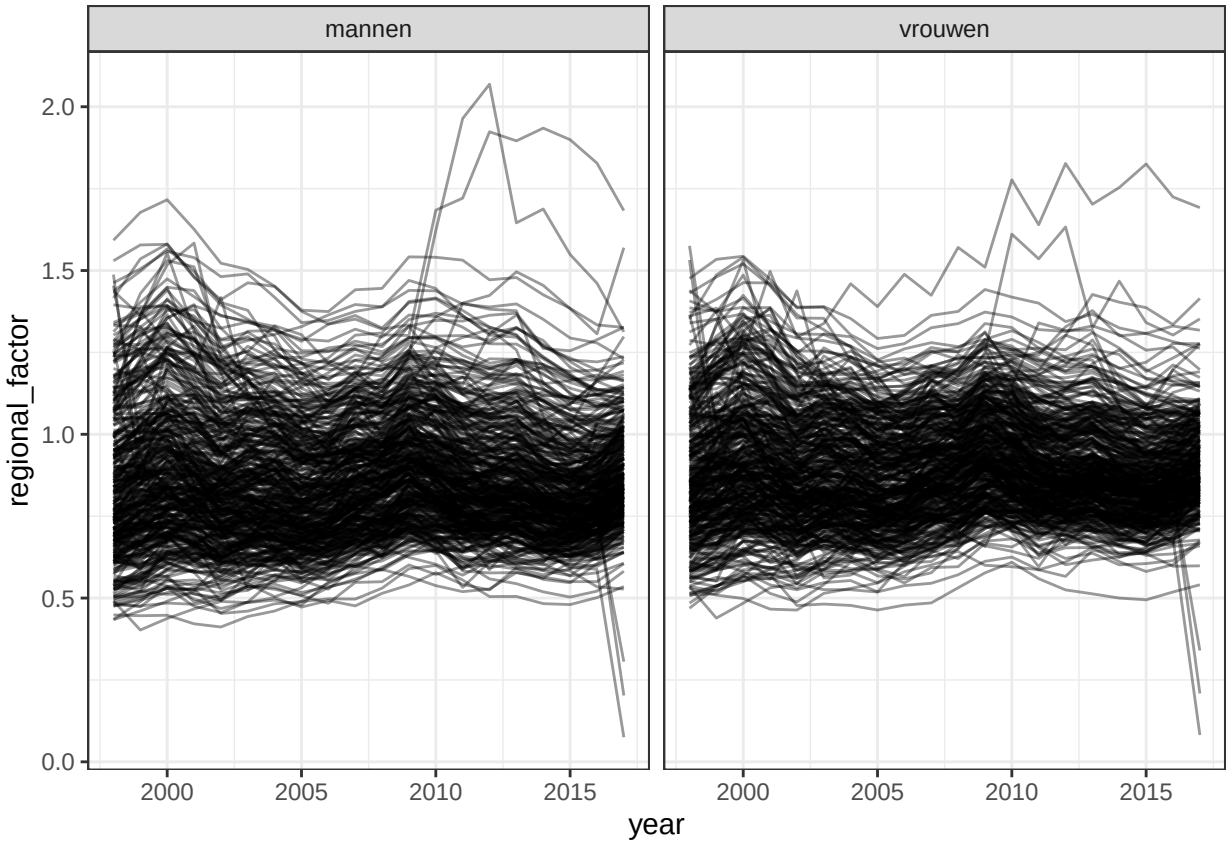

Figure 5: De regionale factor

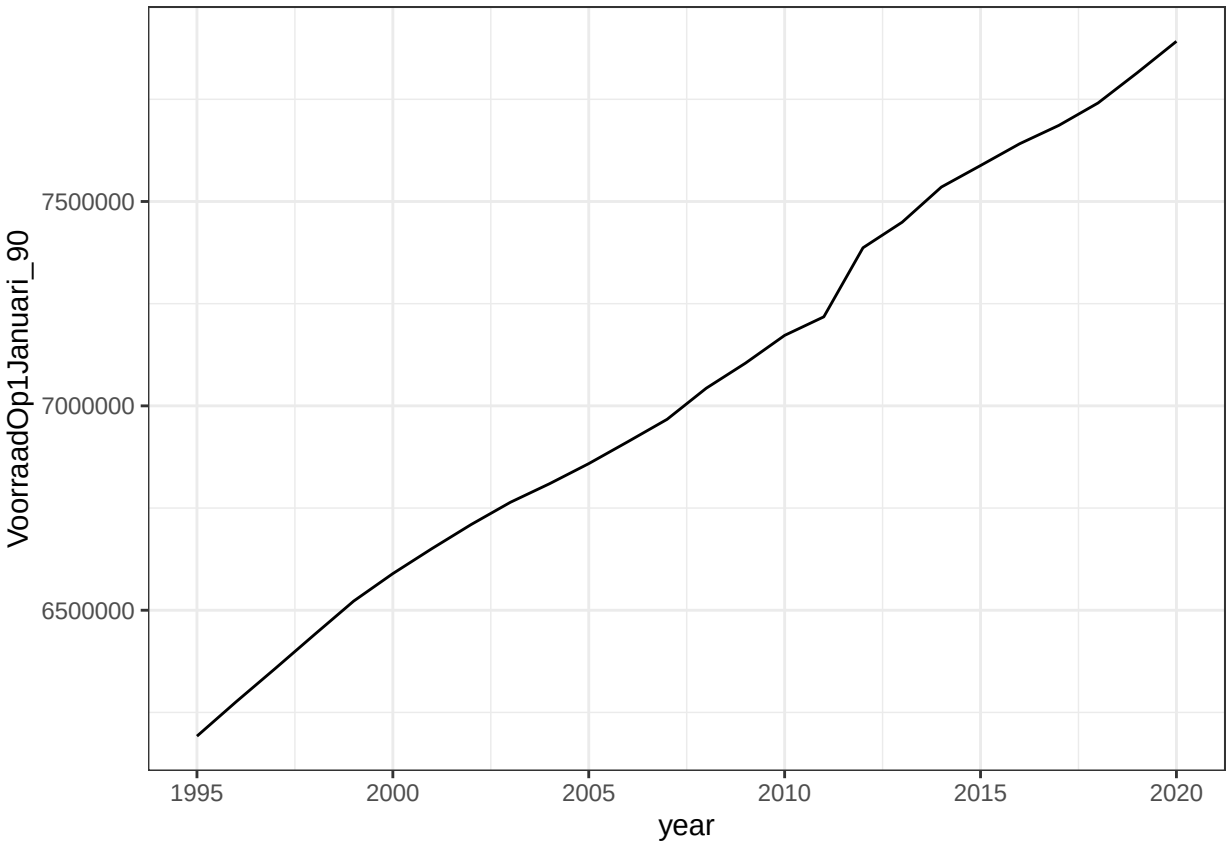

Figure 6: Woningvoorraad (landelijk)

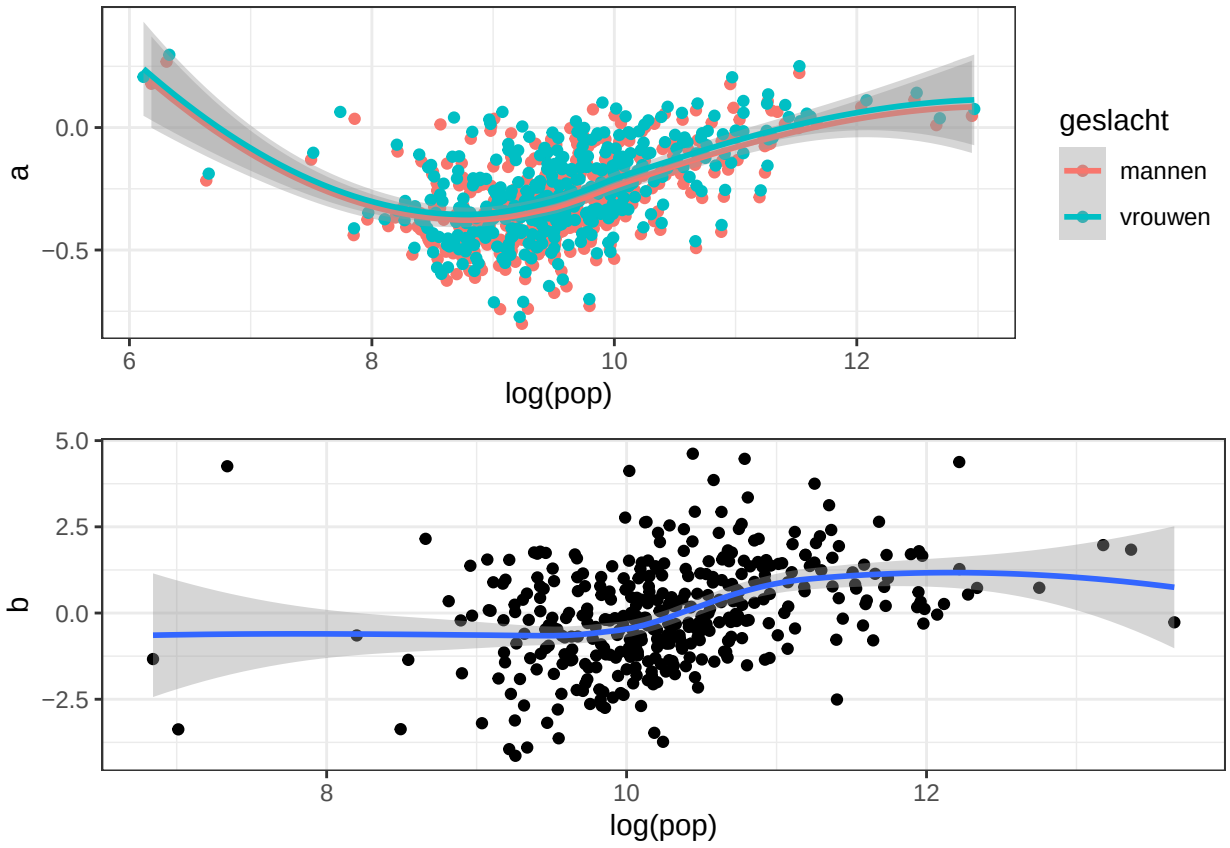

Figure 7: Parameters van het regiomodel: geschatte alpha en beta versus bevolking (log)

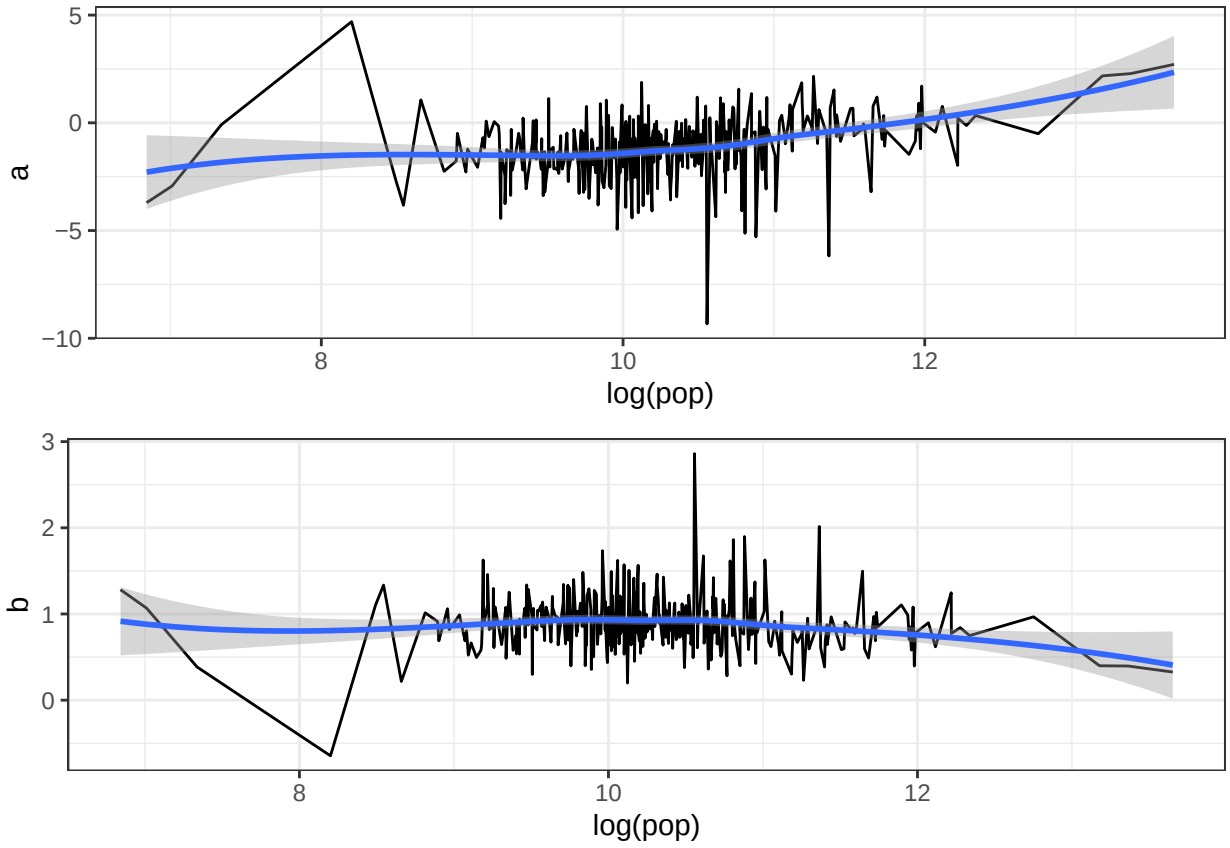

Figure 8: Parameters van het woningmodel: geschatte alpha en beta versus bevolking (log)

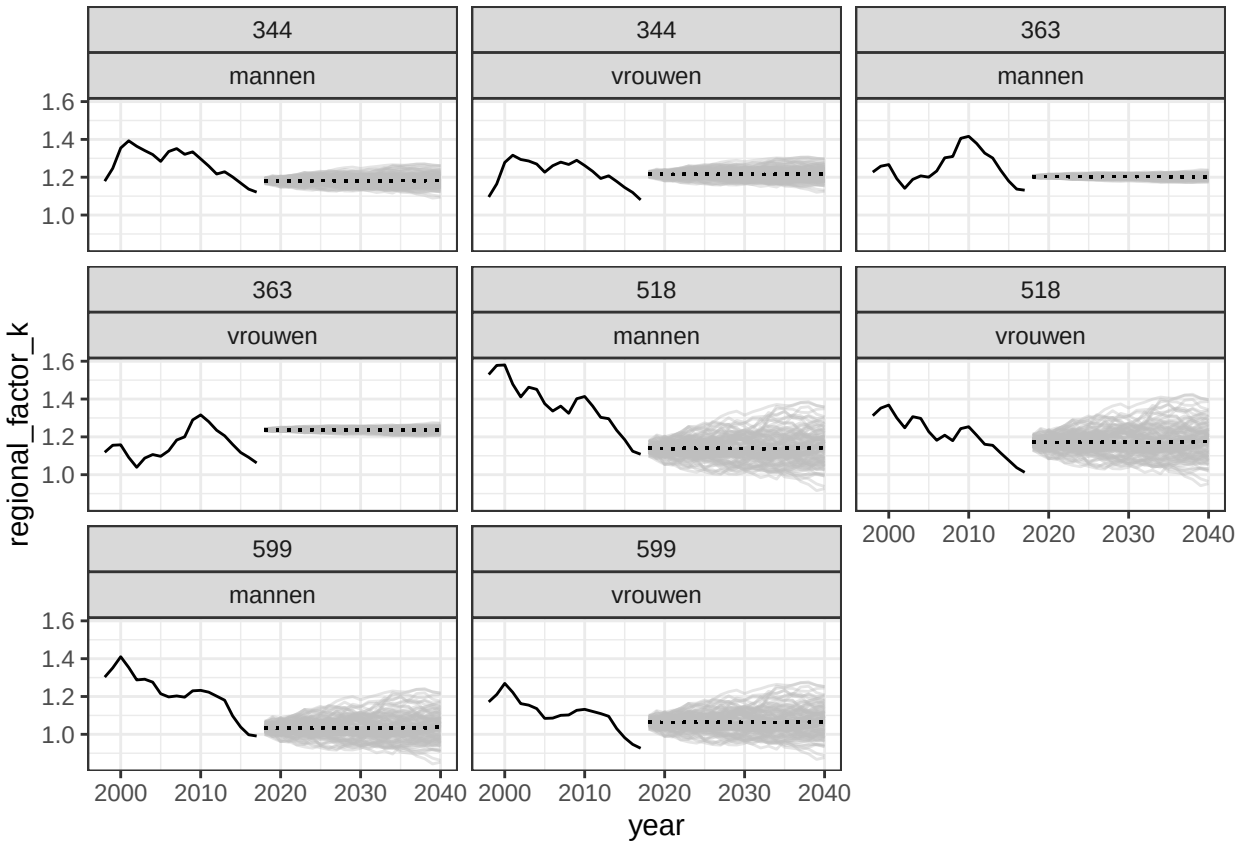

Figure 9: Simulaties regionale factor

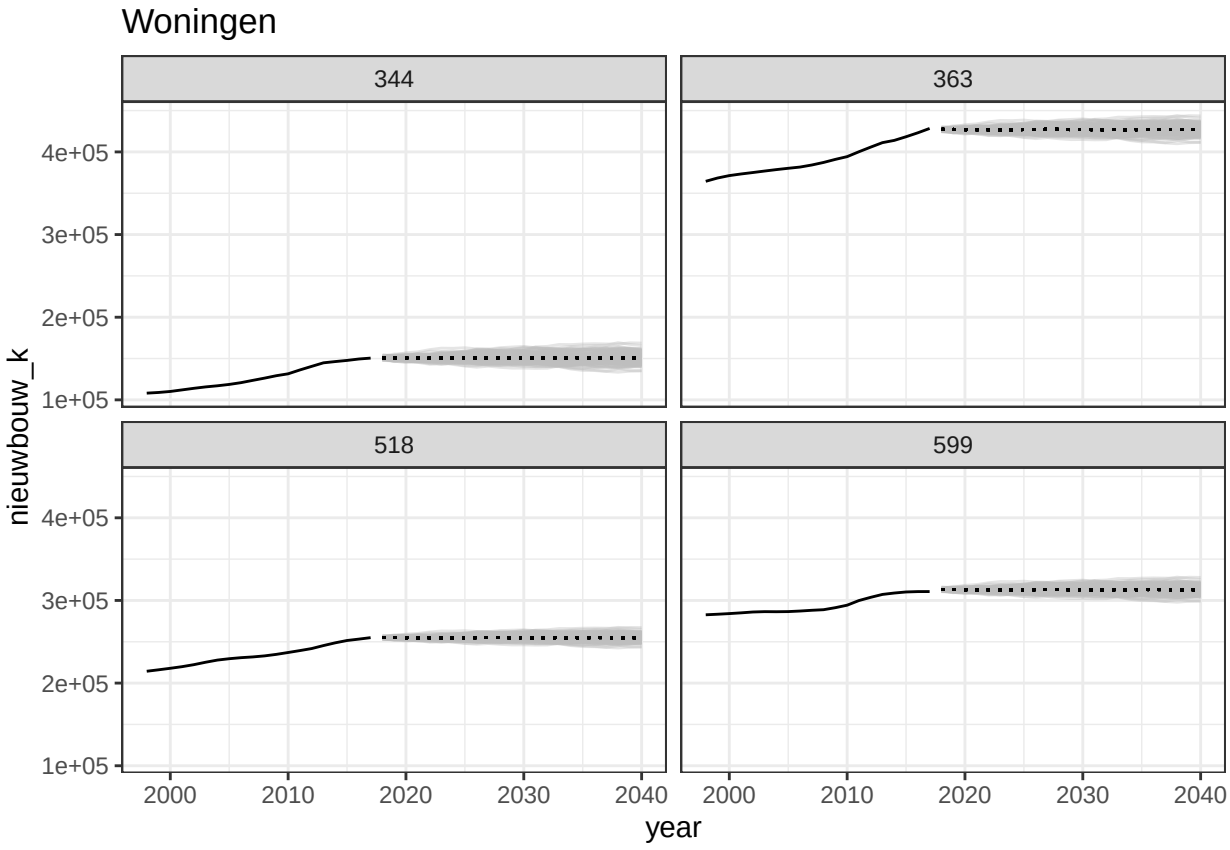

Figure 10: Simulaties woningbouw

Literatuur

Booth, Heather, Rob J Hyndman, Leonie Tickle, and Piet De Jong. 2006. "Lee-Carter Mortality Forecasting: A Multi-Country Comparison of Variants and Extensions." *Demographic Research* 15. JSTOR: 289–310.

Brouhns, Natacha, Michel Denuit, and Jeroen K Vermunt. 2002. "A Poisson Log-Bilinear Regression Approach to the Construction of Projected Lifetables." *Insurance: Mathematics and Economics* 31 (3). Elsevier: 373–93.

Currie, Iain D. 2013. "Fitting Models of Mortality with Generalized Linear and Non-Linear Models." *Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, Eh14 4as, UK*.

Hyndman, Rob J, and Heather Booth. 2008. "Stochastic Population Forecasts Using Functional Data Models for Mortality, Fertility and Migration." *International Journal of Forecasting* 24 (3). Elsevier: 323–42.

Hyndman, Rob J, Anne B Koehler, Ralph D Snyder, and Simone Grose. 2002. "A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods." *International Journal of Forecasting* 18 (3). Elsevier: 439–54.

Keilman, Nico. 2001. "Uncertain Population Forecasts." *Nature* 412 (6846). Nature Publishing Group: 490–91.

Renshaw, Arthur E, and Steve Haberman. 2008. "On Simulation-Based Approaches to Risk Measurement in Mortality with Specific Reference to Poisson Lee–Carter Modelling." *Insurance: Mathematics and Economics*

42 (2). Elsevier: 797–816.

Stoeldraijer, Lenny, Coen van Duin, Leo van Wissen, and Fanny Janssen. 2013. "Impact of Different Mortality Forecasting Methods and Explicit Assumptions on Projected Future Life Expectancy: The Case of the Netherlands." *Demographic Research* 29. JSTOR: 323–54.

Turner, Heather, and David Firth. 2007. "Generalized Nonlinear Models in R: An Overview of the Gnm Package." ESRC National Centre for Research Methods.

Wiśniowski, Arkadiusz, Peter WF Smith, Jakub Bijak, James Raymer, and Jonathan J Forster. 2015. "Bayesian Population Forecasting: Extending the Lee-Carter Method." *Demography* 52 (3). Springer: 1035–59.