Voorspelfouten bij de toepassing van Markov-modellen in de personeelsplanning

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

VAKGROEP KANSREKENING, STATISTIEK OPERATIONS RESEARCH

Voorspelfouten bij de toepassing van Markov-modellen in de personeelsplanning

door

E. van der Beek

Memorandum COSOR 77-13

Eindhoven, juni 1977

Nederland

Samenvatting

Markov-modellen, zoals deze bij de personeelsplanning worden gebruikt, leveren geen exacte voorspellingen voor de toekomstige bezetting van de rangen binnen een te beschouwen organisatie.
Een reden hiervoor is, dat bijvoorbeeld schommelingen in het personeelsverloop nooit exact kunnen worden voorspeld uit beschikbare historische gegevens over het personeelssysteem.
In dit rapport zullen de statistische aspecten van de toepassing van Markov-modellen in de personeelsplanning worden belicht.
Fouten tengevolge onnauwkeurige schattingen van overgangswaarschijnlijkheden zullen worden beschouwd en worden vergeleken met de variantie in de schattingen tengevolge van het stochastische karakter van het model.
Daarnaast zullen schattingen voor de verwachte kwadratische fout in de voorspellingen worden gegeven.
Voorts zullen de gevolgen voor de kwaliteit van de voorspellingen van het clusteren van toestanden in het Markov-model, waarbij binnen één cluster historische promotie- en verloopgegevens worden geaggregeerd, worden besproken.

1. Inleiding

In de literatuur over personeelsplanning treft men veel modellen met een Markov-keten-struktuur aan. Dit rapport is gebaseerd op een gegeneraliseerd Markov-model voor het dynamische gedrag van een individuele werknemer. Het betreft een systeem, FORMASY geheten, ten behoeve van voorspellingen en recrutering in personeelssystemen, ontwikkeld aan de T.H. Eindhoven.
In FORMASY [5] worden de individuele personeeldleden naar een aantal kenmerken ingedeeld in categorieën of toestanden. Als kenmerken zijn in FORMASY onderscheiden: rang, leeftijdsgroep, opleidingsniveau en rangsanciënniteit. De mogelijkheid bestaat om ook andere kenmerken op te nemen. Alle personeeldleden die het systeem verlaten worden in een speciale categorie ingedeeld. Deze categorie zal het verloop worden genoemd.
In dit rapport zal worden verondersteld, dat de individuele personeelsleden een onderling onafhankelijk promotie- en verloopgedrag kennen. Waar deze aanname niet noodzakelijk is,zal dit ter plaatse worden vermeld. Verder wordt aangenomen, dat overgangen tussen twee categorieën geschieden volgens vaste overgangswaarschijnlijkheden, die slechts afhangen van deze categorieën en niet van het overgangstijdstip of het betreffende personeelslid.
Eenvoudigheidshalve zal worden aangenomen, dat gedurende de periode waarin de bezetting wordt voorspeld geen recruteringen plaatsvinden.

In §2 zullen een aantal notationele zaken aan de orde komen en aanvullende veronderstellingen worden gemaakt.
Vervolgens zal in §3 het effect op de verwachte bezetting van een variatie van de overgangswaarschijnlijkheden worden besproken. De diverse foutenbronnen zullen in §4 worden behandeld. Van deze foutenbronnen wordt in §5 de statistische fout nader beschouwd. Daarbij zullen zowel de frequentistische methoden als de Bayesiaanse benadering ter sprake komen.
Tenslotte zal in §7 worden nagegaan wat de effecten zijn van het clusteren van categorieën voor de kwaliteit van de voorspellingen.

2. Vooronderstellingen en notaties

Laat S:= {1,2,...,k+1} de verzameling van alle categorieën zijn, waartoe de personeelsleden kunnen behoren. Hierbij wordt categorie k+1 gevormd door het verloop en is S^*:= {1,2,...,k} de verzameling doorgangstoestanden. Laat nu $\underline{n}_i(t)$ het aantal personeelsleden zijn, dat op tijdstip t behoort tot categorie i. Laat verder $\underline{n}_{ij}(t)$ het aantal leden zijn, dat tussen de tijdstippen t en t+1 overgaat van categorie i naar categorie j. Aangenomen wordt, dat de organisatie een hiërarchische struktuur heeft,zodat $\underline{n}_{ij}(t) = 0$ als $j \leq i$ $(i,j \in S^*)$.
In FORMASY is aan deze aanname voldaan. De kenmerken rang en rangsancienniteit bewerkstellingen dit.
Tussen de bezetting op tijdstip t+1 en de personeelsstromen tussen de tijdstippen t en t+1 bestaat de volgende relatie:

$$\underline{n}_j(t+1) = \sum_{i=1}^{k+1} \underline{n}_{ij}(t) \qquad (j \in S).$$

Laat voor ieder lid de kans op promotie van categorie i naar categorie j gegeven worden door p_{ij}.
Hierbij is verondersteld dat $0 \leq p_{ij} \leq 1$, $\sum_{j=1}^{k+1} p_{ij} = 1$, $(i,j \in S)$ en tevens dat $p_{ij} = 0$ $(j \leq 1,\ i,j \in S)$.

Laat $P = \{p_{ij}\}$ de overgangsmatrix zijn en laat de bezettingsvector gegeven worden door $\underline{n}(t) = (\underline{n}_1(t),\ldots,\underline{n}_{k+1}(t))$, dan geldt voor de verwachte bezetting op tijdstip t+1, gegeven dat $\underline{n}(t) = n(t)$:

$$\mathbb{E}\,[\underline{n}(t+1) \mid \underline{n}(t) = n(t)] = n(t)P \qquad (1)$$

In het algemeen is het juist deze grootheid, die men als voorspelling voor de bezetting gebruikt.
Nemen we nu aan dat de bezetting op tijdstip 0 bekend is en gegeven wordt door n(0), dan geldt voor de verwachte bezetting na t perioden:

$$\mathbb{E}\,\underline{n}(t) = n(0)P^t \qquad (2)$$

P is zelden een bekende matrix. Een schatting voor P kan worden verkregen door te inventariseren welke overgangen tijdens een waarnemingsperiode zijn gemaakt. De aldus verkregen schatting $\hat{P}$ levert, gesubstitueerd in (2), de gevolgen van een in de toekomst ongewijzigd promotiebeleid voor de personeelsbezetting. Met (2) kan eveneens worden nagegaan, wat het effect is van een nieuw te hanteren promotiestrategie. Opgemerkt kan nog worden, dat in deze paragraaf de aanname van een onafhankelijk promotiegedrag niet is gebruikt.

3. Variatie van overgangspercentages

3.0. Inleiding

In deze paragraaf zal steeds worden aangenomen, dat de overgangsmatrix P bekend is.
Het doel is na te gaan, wat het effect is op de verwachte bezetting van variatie van overgangswaarschijnlijkheden d.w.z. van het hanteren van een gewijzigde promotiestrategie.
In §3.1 zal een overzicht worden gegeven van het noodzakelijke gereedschap uit de matrixtheorie. Vervolgens zal in §3.2 het verschil tussen twee promotiestrategieën, gerepresenteerd door Markov-matrices P_1 en P_2, worden beschouwd. De resultaten uit deze deelparagraaf zijn bruikbaar bij de analyse van het clusteren van toestanden. Binnen één cluster worden voor iedere categorie dezelfde overgangspercentages gehanteerd, welke door weging uit de afzonderlijke overgangspercentages worden berekend.
Het clusteren kan daarom worden beschouwd als het op een gerichte wijze variëren van overgangspercentages.

3.1. Normen en matrices

De lezer, die vertrouwd is met de matrixrekening zij direct verwezen naar §3.2. De lemma's welke worden vermeld zullen niet bewezen worden. De bewijzen zijn echter steeds elementair.

Laat $A = \{a_{ij}\} \in M_{m,n}$ zijn, d.w.z. laat A een matrix zijn met m rijen en n kolommen.
Voor deze matrix worden de volgende normen gedefinieerd:

Definitie 3.1.: $\|A\|_1 := \sum_{i=1}^{m} \sum_{j=1}^{m} |a_{ij}|$

$$\|A\|_\infty := \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$$

Nu is gemakkelijk na te gaan, dat $\|\cdot\|_1$ en $\|\cdot\|_\infty$ inderdaad matrix-normen zijn.

Lemma 3.1: Zowel $\|\cdot\|_1$ als $\|\cdot\|_\infty$ zijn submultiplicatief, d.w.z. als $A = \{a_{ij}\} \in M_{m,k}$ en $B = \{b_{ij}\} \in M_{k,n}$ dan geldt:

i) $\|AB\|_1 \leq \|A\|_1 \|B\|_\infty \leq \|A\|_1 \|B\|_1$

ii) $\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty$.

Definitie 3.2: $D = \{d_{ij}\} \in M_{n,n}$ heet diagonaalmatrix als geldt, dat $d_{ij} = 0$ $(i \neq j)$

Een diagonaalmatrix $D \in M_{n,n}$ wordt als $D = \text{diag}(d_1,\ldots,d_n)$ genoteerd.

Definitie 3.3: $A \in M_{m,n}$ heet positief wanneer:

$$\forall_{1\leq i\leq m} \; \forall_{1\leq j\leq n} : a_{ij} \geq 0$$

We noteren dit als $A \geq 0$ en noemen $A \geq B$ wanneer $A - B \geq 0$

Definitie 3.4: $A \in M_{m,n}$ heet strikt positief wanneer:

$$\forall_{1\leq i\leq m} \; \forall_{1\leq j\leq n} : a_{ij} > 0$$

We noteren dit als $A > 0$.

Opmerking: $A < 0$, $A \leq 0$ worden op analoge wijze gedefinieerd.

Lemma 3.2: $\geq$, $\leq$, $>$, $<$ zijn partiële ordeningsrelaties op $M_{m,n}$

Definitie 3.5: Voor $A,B \in M_{m,n}$ worden het maximum van A en B en het minimum van A en B gedefinieerd door:

$$\max(A,B) := \{\max(a_{ij},b_{ij})\}$$

$$\min(A,B) := \{\min(a_{ij},b_{ij})\}$$

Lemma 3.3: Voor $A,B \in M_{m,n}$ geldt:

i) $\max(A,B) \geq A$, $\max(A,B) \geq B$

ii) $\min(A,B) \leq A$, $\min(A,B) \leq B$

iii) $\max(A,B) + \min(A,B) = A + B$

Lemma 3.4: Voor $A,B \in M_{m,n}$, $A \geq 0$, $B \geq 0$, $A \geq B$ geldt:

$$\|A\|_1 = \|B\|_1 + \|A-B\|_1$$

$$\|A\|_\infty \leq \|B\|_\infty + \|A-B\|_\infty$$

Lemma 3.5: Als $A, B, C \in M_{m,n}$ $B \geq 0$, $C \leq 0$ en $C \leq A \leq B$, dan geldt:

$$\|A\|_1 \leq \|B-C\|_1 = \|B\|_1 + \|C\|_1$$
$$\|A\|_\infty \leq \|B-C\|_\infty \leq \|B\|_\infty + \|C\|_\infty$$

Laat $P \in M_{n,n}$ een Markov-matrix d.w.z.

$$P = \{p_{ij}\} \geq 0 \wedge \forall_{1\leq i\leq n} : \sum_{j=1}^{n} p_{ij} = 1$$

dan is $\|P\|_1 = n, \quad \|P\|_\infty = 1$

Lemma 3.6: Voor $A \in M_{m,n}$, $B \in M_{n,k}$ $P \in M_{n,n}$ met P een Markov-matrix geldt:

$$\|AP\|_1 \leq \|A\|_1 \|P\|_\infty = \|A\|_1$$
$$\|AP\|_\infty \leq \|A\|_\infty$$
$$\|PB\|_\infty \leq \|B\|_\infty$$

Opmerking: Het is i.h.a. niet zo dat geldt:

$$\|PB\|_1 \leq \|B\|_1$$

3.2. Bovengrenzen voor de variatie in de bezettingsvector n(t) ten gevolge van variatie van de overgangswaarschijnlijkheden.

Laat P_1 en $P_2 \in M_{n,n}$ zijn, met P_1 en P_2 beide Markov-matrices. Laat Δ_1 en Δ_2 gedefinieerd worden, door:

$$\Delta_1 := P_1 - \min(P_1, P_2)$$
$$\Delta_2 := P_2 - \min(P_1, P_2)$$

dan geldt:

i) $\Delta_1 \geq 0, \quad \Delta_2 \geq 0$

ii) $\forall_{1\leq i\leq n} : \sum_{j=1}^{n} \{\Delta_1\}_{ij} = \sum_{j=1}^{n} \{\Delta_2\}_{ij} =: \delta_i$

Verder is:

$$\|\Delta_1\|_\infty = \|\Delta_2\|_\infty = \max_{1\leq i\leq n} \delta_i =: \delta.$$

en

$$\|\Delta_1\|_1 = \|\Delta_2\|_1 = \sum_{i=1}^{n} \delta_i$$

We kunnen nu de volgende stelling formuleren:

<u>Stelling 3.1</u>: $\|P_1^k - P_2^k\|_1 \leq 2 \left(\sum_{i=1}^{n} (1-(1-\delta_i)(1-\delta)^{k-1}) \right)$

$$\|P_1^k - P_2^k\|_\infty \leq 2(1-(1-\delta)^k)$$

<u>Bewijs:</u> Uit lemma 3.3 volgt:

$$[\min(P_1,P_2)]^k - P_2^k \leq P_1^k - P_2^k \leq P_1^k - [\min(P_1,P_2)]^k$$

We definiëren (onder de veronderstelling dat $\delta < 1$)

$$D := \text{diag} \left(\frac{1-\delta_1}{1-\delta}, \ldots, \frac{1-\delta_n}{1-\delta} \right)$$

Daar $\delta_i \leq \delta$ $(i=1,\ldots,n)$ volgt, dat $D \geq I$.

Voor

$$Q := D^{-1} . \min(P_1,P_2)$$

geldt:

i) $Q := \{q_{ij}\} \geq 0$

ii) $\forall_{1\leq i\leq n} : \sum_{j=1}^{n} q_{ij} = \frac{1-\delta}{1-\delta_i}(1-\delta_i) = 1-\delta$

Nu gelden de volgende ongelijkheden:

$$P_1^k - P_2^k \leq P_1^k - (DQ)^k \leq P_1^k - DQ^k$$

$$P_2^k - P_1^k \leq P_2^k - (DQ)^k \leq P_2^k - DQ^k$$

Daar $P_1^k \geq D.Q^k$ en $P_2^k \geq D.Q^k$ volgt uit lemma 3.5:

$$\|P_1^k - P_2^k\|_1 \leq \|P_1^k + P_2^k - 2DQ^k\|_1 =$$

$$= 2n - 2 \sum_{i=1}^{n} \frac{1-\delta_i}{1-\delta} (1-\delta)^k =$$

$$= 2\left(\sum_{i=1}^{n} (1-(1-\delta_i)(1-\delta)^{k-1}) \right).$$

en ook:

$$\|P_1^k - P_2^k\|_\infty \leq \|P_1^k \quad P_2^k - 2DQ^k\|_\infty$$

$$= 2 - 2(1-\delta)^k =$$

$$= 2(1-(1-\delta)^k)$$ □

Laat $n_1(k) := n(0)\ P_1^k$ en $n_2(k) := n(0)P_2^k$ zijn.

Voor $v(k) := n_1(k) - n_2(k)$ kan de volgende afschatting worden gegeven:

Stelling 3.2: $\|v(k)\|_1 \leq \|n(0)\|_1\ 2(1-(1-\delta)^k)$

Bewijs: $\|v(k)\|_1 = \|n(0)\ (P_1^k - P_2^k)\|_1 =$

$$\leq \|n(0)\|_1\ \|P_1^k - P_2^k\|_\infty$$

$$= \|n(0)\|_1\ 2(1-(1-\delta)^k)$$ □

Opmerking: Uit stelling 3.2 volgt dat in eerste orde benadering geldt:

$$\|v(k)\|_1 = \|n_1(k) - n_2(k)\|_1 \leq 2k\delta\ \|n(0)\|_1$$

4. Foutenbronnen

Laat $\hat{P}$ een overgangsmatrix zijn, die geschat is uit historische gegevens over een waarnemingsperiode. Dit betekent, dat $\hat{P}$ vastgelegd is door het stochastische karakter van het model in deze waarnemingsperiode.
De voorspelde bezetting op tijdstip t gebaseerd op deze schatting wordt nu:

$$\hat{n}(t) = n(0)\hat{P}^t \qquad (2^a)$$

Bij de voorspelling van $\underline{n}(t)$ wordt daarmee een fout gemaakt, gelijk aan:

$$\underline{n}(t) - \hat{n}(t) = \underline{n}(t) - n(0)\hat{P}^t.$$

Deze fout kan als volgt worden gesplitst:

$$\underline{n}(t) - \hat{n}(t) = (\underline{n}(t) - \mathbb{E}\,\underline{n}(t)) + (\mathbb{E}\,\underline{n}(t) - n(0)\hat{P}^t)$$

$$= (\underline{n}(t) - n(0)P^t) + n(0)(P^t - \hat{p}^t) \qquad (3)$$

De verschilvector $n(t)-n(0)P^t$ treedt op ten gevolge van de <u>statistische fout</u>. Deze fout wordt gemaakt bij het voorspellen, wanneer P bekend is. De tweede fout welke wordt gemaakt en die wordt gegeven door $n(0)(P^t-\hat{P}^t)$ wordt de <u>schattingsfout</u> genoemd. Deze schattingsfout ligt door de bepaling van $\hat{P}$ op tijdstip 0 weliswaar vast, doch is onbekend omdat P onbekend is. Om de verwachte kwadratische schattingsfout te schatten zullen in §6 een tweetal methoden worden gehanteerd nl. de Bayesiaanse schattingsmethode en een frequentische methode. Een derde foutenbron, die kan worden onderscheiden is de <u>specificatiefout</u>. Deze fout treedt op als niet exact aan de modelaannamen is voldaan.
Een oorzaak hiervan kan zijn, dat geen goede indeling in categorieën is gekozen. Als door een verdere verfijning van de indeling in categorieën beter aan de modelveronderstellingen kan worden voldaan,wordt gesproken van de z.g. aggregatiefout.
Het kan echter ook noodzakelijk zijn, dat het model zo wordt aangepast, dat de overgangswaarschijnlijkheden worden opgevat als stochastische variabelen met verschillende realisaties op de diverse tijdstippen.
Bartholomew [3] bespreekt deze laatste vorm van de specificatiefout in samenhang met de statistische fout en de schattingsfout. In dit rapport zullen statistische fout en aggregatiefout in hun onderlinge samenhang worden besproken.

5. De statistische fout

Aangenomen wordt in deze paragraaf, dat de overgangsmatrix P bekend is. Het doel is inzicht te krijgen in de statistische fout, gegeven door $n(t) - n(0)P^t$. Voor de verwachte kwadratische fout $M_{ST}(t)$ in de geschatte bezetting op tijdtip t, volgt m.b.v. (2):

$$M_{ST}(t) = \mathbb{E}\,[\underline{n}(t) - n(0)P^t]'\,[\underline{n}(t) - n(0)P^t]$$

$$= \mathbb{E}[\underline{n}(t) - \mathbb{E}\,\underline{N}(t)]'\,[\underline{n}(t) - \mathbb{E}\,\underline{n}(t)]. \qquad (4)$$

Nu blijkt, dat $M_{ST}(t)$, de te hanteren maat voor de kwaliteit van de voorspelling op tijdstip, juist de variantie-covariantie matrix is van de componenten van $\underline{n}(t)$, wanneer P bekend is.
Bartolomew [2] heeft de volgende recurrente betrekkingen afgeleid tussen de elementen van $M_{ST}(t+1)$ en $M_{ST}(t)$:

$$\operatorname{cov}(\underline{n}_i(t+1), \underline{n}_j(t+1)) = \sum_{\ell=1}^{k+1}\sum_{r=1}^{k+1} p_{\ell i}p_{rj}\operatorname{cov}(\underline{n}_\ell(t), \underline{n}_r(t)) +$$

$$+ \sum_{\ell=1}^{k+1} (\delta_{ij}-p_{\ell j})p_{\ell i}\,\mathbb{E}\,\underline{n}_\ell(t). \qquad (5)$$

Omdat de bezettingsvector n(0) bekend is, kan (5) voor t = 0 als volgt worden vereenvoudigd:

$$\operatorname{cov}(\underline{n}_i(1)\underline{n}_j(1)) = \sum_{\ell=1}^{k+1} (\delta_{ij}-p_{\ell j})p_{\ell i}\,n_\ell(0) \qquad (6)$$

De covarianties voor de tijdstippen t = 1,2,... kunnen nu iteratief worden berekend.

6. De schattingsfout

Zoals reeds is opgemerkt, kan de schattingsfout vanuit een tweetal gezichtspunten worden benaderd nl. volgens frequentistische methoden, zie §6.1 m.b.v. een Bayesiaanse aanpak, zie §6.3. In §6.2 wordt de frequentistische methode ook toegepast bij de voorspelling over in meerdere perioden.

6.1. De frequentistische methode

Vaak wordt P geschat op grond van promotiestromen uit het verleden. Stel nu, dat gedurende de perioden [-T, -T+1], [-T+1, -T+2],..., [-1,0] het systeem wordt waargenomen.
Laat $\underline{N}_i(-\tau)$ $(\tau=1,\ldots,T)$ de bezetting zijn van categorie i op tijdstip $-\tau$ en laat verder $\underline{N}_{ij}(-\tau)$ de fractie van de bezetting van categorie i zijn, welke naar categorie j overgaat gedurende het tijdsinterval $(-\tau,-\tau+1)$ $(t=1,\ldots,T)$.
Geaggregeerd stelt $\underline{N}_i := \sum_{\tau=1}^{T} \underline{N}_i(-\tau)$ de totale bezetting van categorie i gedurende de waarnemingsperiode voor en is verder $\underline{N}_{ij} := \sum_{\tau=1}^{T} \underline{N}_{ij}(-\tau)$ de fractie van deze bezetting, die gedurende de waarnemingsperiode overgaat naar categorie j.
Er geldt nu:

$$\underline{N}_i = \sum_{j=1}^{k+1} \underline{N}_{ij} \tag{7}$$

Eenvoudig is na te gaan, dat de maximum-likelihoodschatter $\hat{\underline{p}}_{ij}$ voor p_{ij} gegeven wordt door:

$$\hat{\underline{p}}_{ij} = \frac{\underline{N}_{ij}}{\underline{N}_i} \tag{8}$$

Verondersteld wordt, dat $\hat{\underline{P}} := \{\hat{\underline{P}}_{ij}\}$ geschat wordt uit een stochastisch proces onafhankelijk van het stochastische proces, dat de personeelsstromen in de toekomst beschrijft.
M.a.w. de stochastische processen $\{\underline{N}_i(-\tau) \mid t \in \mathbb{N}\}$ en $\{\underline{n}_j(t) \mid t \in \mathbb{N} \cup \{0\}\}$ $(i,j \in S)$ worden onafhankelijk verondersteld.
In de praktijk zullen $\underline{n}(0)$ en $\hat{\underline{P}}$ i.h.a. niet onafhankelijk zijn. Aan het slot van deze deelparagraaf zal de consequentie van deze aanname voor de kwaliteit van de voorspellingen aan de orde komen.
Onze aandacht richt zich op de invloed van de schattingsfout op de kwaliteit van de voorspellingen.
Volgens (2^a) geldt voor de geschatte bezetting op tijdstip 1 wanneer n(0) is gegeven:

$$\underline{\hat{n}}(1) = n(0)\,\underline{\hat{P}} \tag{9}$$

Voor de verwachte kwadratische fout M(1) van de voorspelde bezetting na één periode wordt de volgende uitdrukking gevonden:

$$\begin{aligned} M(1) &:= \mathbb{E}[\underline{n}(1)-\underline{\hat{n}}(1)]'\,[\underline{n}(1)-\underline{\hat{n}}(1)] \\ &= \mathbb{E}[\underline{n}(1)-\mathbb{E}\,\underline{n}(1)]'\,[\underline{n}(1)-\mathbb{E}\,\underline{n}(1)]+ \\ &+ \mathbb{E}[(\mathbb{E}\,\underline{n}(1)-\underline{\hat{n}}(1))'\,(\mathbb{E}\,\underline{n}(1)-\underline{\hat{n}}(1)] \\ &= M_{ST}(1) + M_E(1) \end{aligned} \tag{10}$$

waarbij $M_E(1) := \mathbb{E}[(\mathbb{E}\ \underline{n}(1)-\underline{\hat{n}}(1)'\ (\mathbb{E}\ \underline{n}(1)-\underline{\hat{n}}(1)]$ een maat is voor de geïntroduceerde onzuiverheid.
Er geldt:

$$\begin{aligned} M_E(1) &= \mathbb{E}[(n(0)(P-\underline{\hat{P}}))'\ (n(0)(P-\underline{\hat{P}}))] \\ &= \mathbb{E}[(\underline{\hat{P}}-P)'\ n(0)'n(0)(\underline{\hat{P}}-P)] \end{aligned} \tag{11}$$

Componentsgewijze is:

$$\{M_E\ 1)\}_{j\ell} = \sum_{i_1=1}^{k+1}\ \sum_{i_2=1}^{k+1} n_{i_1}(0)n_{i_2}(0)\ \mathrm{cov}(\underline{\hat{P}}_{i_1 j},\underline{\hat{P}}_{i_2 \ell})$$

Vanwege de aanname dat het promotiegedrag voor verschillende individuen onafhankelijk is en omdat ieder individu maximaal één bezoek brengt aan een bepaalde categorie vinden we, dat $\mathrm{cov}(\underline{\hat{P}}_{i_1 j},\underline{\hat{P}}_{i_2 \ell}) = 0 \quad (i_1 \neq i_2)$.
Daardoor geldt:

$$\{M_E(1)\}_{j\ell} = \sum_{i=1}^{k+1} n_i^2(0)\ \mathrm{cov}(\underline{\hat{P}}_{ij},\underline{\hat{P}}_{i\ell}) \tag{12}$$

In het vervolg van deze deelparagraaf worden eerst uitdrukkingen afgeleid om $\mathrm{cov}(\underline{\hat{P}}_{ij},\underline{\hat{P}}_{i\ell})$ te kunnen schatten. Vervolgens zullen de relatieve bijdragen van statistische- en schattingsfout tot de verwachte kwadratische voorspelfout vergeleken worden. Tenslotte volgt nog een opmerking over de afhankelijkheid van $\underline{n}(0)$ en $\underline{\hat{P}}$.

Lemma 6.1:

Voor $\hat{p}_{ij}(-\tau) := \frac{N_{ij}(-\tau)}{N_i(-\tau)}$ en $\hat{p}_{i\ell}(-\tau) := \frac{\hat{N}_{i\ell}(-\tau)}{N_i(-\tau)}$ $\quad (i,j,\ell \in S,\ j \neq \ell)$

geldt:

$$\text{i) } \operatorname{var}(\hat{p}_{ij}(-\tau) \mid N_i(-\tau)) = N_i(-\tau) = \frac{1}{N_i(-\tau)}\, p_{ij}(1-p_{ij}) \tag{13}$$

$$\operatorname{cov}(\hat{p}_{ij}(-\tau)\ \hat{p}_{i\ell}(-\tau)) \mid N_i(-\tau) = N_i(-\tau)) = -\frac{1}{N_i(-\tau)}\, p_{ij}p_{i\ell} \tag{14}$$

$$\text{ii) } \operatorname{var}(\hat{p}_{ij}(-\tau)) = \mathbb{E}\, \frac{1}{N_i(-\tau)}\, p_{ij}(1-p_{ij}) \tag{13a}$$

$$\operatorname{cov}(\hat{p}_{ij}(-\tau),\ \hat{p}_{i\ell}(-\tau)) = -\,\mathbb{E}\, \frac{1}{N_i(-\tau)}\, p_{ij}p_{i\ell} \tag{14a}$$

bewijs: Als geldt, dat $N_i(-\tau) = N_i(-\tau)$ wordt in categorie i a.h.w. een multinomiaal experiment uitgevoerd met parameters $N_i(-\tau)$ en $p_{i1},\ldots,p_{i,k+1}$.
Het bewijs van i) volgt nu uit:

$$\operatorname{var}(\hat{p}_{ij}(-\tau) \mid N_i(-\tau) = N_i(-\tau)) =$$

$$= \left(\frac{1}{N_i(-\tau)}\right)^2 \cdot \operatorname{var} N_{ij}(-\tau) =$$

$$= \frac{1}{N_i(-\tau)}\, p_{ij}\ (1-p_{ij}).$$

Analoog volgt, dat:

$$\operatorname{cov}(\hat{p}_{ij}(-\tau),\ \hat{p}_{i\ell}(-\tau)) \mid N_i(-\tau)) =$$

$$= \left(\frac{1}{N_i(-\tau)}\right)^2 \cdot \operatorname{cov}(N_{ij}(-\tau),\ N_{i\ell}(-\tau)) =$$

$$= -\,\frac{1}{N_i(-\tau)}\ p_{ij}p_{i\ell}$$

Voorts geldt, dat:

$$\mathbb{E}\,[p_{ij}(-\tau) \mid N_i(-\tau) = N_i(-\tau)] = p_{ij}$$

zodat nu met behulp van de volgende eigenschap:

$$\sigma^2(\underline{x}) = \mathbb{E}_y\, \sigma_x^2(\underline{x}|\underline{y}) + \sigma_y^2(\mathbb{E}_x(\underline{x}|\underline{y}))$$

uit i) volgt, dat:

$$\mathrm{var}(\hat{\underline{p}}_{ij}(-\tau)) = \mathbb{E}\, \frac{1}{\underline{N}_i(-\tau)}\, p_{ij}(1-p_{ij}) + \sigma^2_{\underline{N}_i(-\tau)}(p_{ij})$$

$$= \mathbb{E}\, \frac{1}{\underline{N}_i(-\tau)}\, p_{ij}(1-p_{ij})$$

Uit het feit, dat:

$$\sigma(\underline{x}_1,\underline{x}_2) = \mathbb{E}_y\, \sigma_{x_1,x_2}((\underline{x}_1,\underline{x}_2) \mid \underline{y}) + \sigma_y\, \mathbb{E}_{x_2}(\underline{x}_2|\underline{y}))$$

volgt verder uit i)

$$\mathrm{cov}(\hat{\underline{p}}_{ij}(-\tau),\hat{\underline{p}}_{i\ell}(-\tau)) = -\,\mathbb{E}\, \frac{1}{N_i(-\tau)}\, p_{ij}p_{i\ell} + \sigma_{\underline{N}_i(-\tau)}\, (p_{ij},p_{i\ell})$$

$$= -\,\mathbb{E}\, \frac{1}{N_i(-\tau)}\, p_{ij}p_{i\ell} \qquad \square$$

N.B. De correctheid van de formules voor $\sigma^2(\underline{x})$ en $\sigma(\underline{x}_1,\underline{x}_2)$ is gemakkelijk in te zien.
Immers:

$$\mathbb{E}_y\, \sigma_x^2(\underline{x}|\underline{y}) + \sigma_y^2\, (\mathbb{E}_x(\underline{x}|\underline{y})) =$$

$$= [\mathbb{E}_y(\mathbb{E}_x(\underline{x}^2|\underline{y})) - \mathbb{E}_y(\mathbb{E}_x(\underline{x}|\underline{y}))^2] + \mathbb{E}_y(\mathbb{E}_x(\underline{x}|\underline{y}))^2$$

$$- (\mathbb{E}_y(\mathbb{E}_x(\underline{x}|\underline{y})))^2 =$$

$$= \mathbb{E}_y(\mathbb{E}_x(\underline{x}^2|\underline{y})) - (\mathbb{E}_y\, \mathbb{E}_x(\underline{x}|\underline{y}))^2 =$$

$$= \mathbb{E}\underline{x}^2 - (\mathbb{E}\underline{x})^2 = \sigma^2(\underline{x})$$

En tevens:

$$\mathbb{E}_y(\sigma_{x_1,x_2}\, (\underline{x}_1,\underline{x}_2\, |\underline{y})) + \sigma_y^2\, (\mathbb{E}_{x_1}(\underline{x}_1|\underline{y}),\, \mathbb{E}_{\underline{x}_2}(\underline{x}_2|\underline{y})) =$$

$$= \mathbb{E}_y(\mathbb{E}_{x_1x_2}(\underline{x}_1,\underline{x}_2|\underline{y})) - \mathbb{E}_y\, (\mathbb{E}_{x_1}(\underline{x}_1|\underline{y})\mathbb{E}_{x_2}(\underline{x}_2|\underline{y}))$$

$$+ \mathbb{E}_y(\mathbb{E}_{x_1}(\underline{x}_1|\underline{y}).\mathbb{E}_{x_2}(\underline{x}_2|\underline{y})) - \mathbb{E}_y\mathbb{E}_{x_1}(\underline{x}_1|\underline{y}).\mathbb{E}_y\mathbb{E}_{x_2}(\underline{x}_2|\underline{y}) =$$

$$= \mathbb{E}_y(\mathbb{E}_{x_1x_2}(\underline{x}_1\underline{x}_2|\underline{y})) - \mathbb{E}_y(\mathbb{E}_{x_1}(\underline{x}_1|\underline{y})).\mathbb{E}_{x_2}(\underline{x}_2|\underline{y}))$$

$$= \mathbb{E}\underline{x}_1\underline{x}_2 - \mathbb{E}\underline{x}_1\mathbb{E}\underline{x}_2 = \sigma(\underline{x}_1,\underline{x}_2)$$

De conclusies uit lemma 6.1 kunnen worden gegeneraliseerd voor schattingen over gegevens uit meerdere perioden.
Er geldt:

Lemma 6.2:

$$\text{Voor } \hat{\underline{p}}_{ij} = \frac{\underline{N}_{ij}}{\underline{N}_i} \text{ en } \hat{\underline{p}}_{i\ell} = \frac{\underline{N}_{i\ell}}{\underline{N}_\ell} \quad (\ell \neq j),\ (i,j,\ell \in S),$$

$$\text{waarbij } \underline{N}_{ij} := \sum_{t=1}^{T} \underline{N}_{ij}(-\tau) \text{ en } \underline{N}_i := \sum_{i=1}^{T} \underline{N}_i(-\tau),$$

geldt:

$$\text{i) } \operatorname{var}(\hat{\underline{p}}_{ij} \mid \underline{N}_i = N_i) = \frac{1}{N_i} p_{ij}(1-p_{ij}) \tag{15}$$

$$\operatorname{cov}(\hat{\underline{p}}_{ij}\hat{\underline{p}}_{i\ell} \mid \underline{N}_i = N_i) = -\frac{1}{N_i} p_{ij}p_{i\ell} \tag{16}$$

$$\text{ii) } \operatorname{var}(\hat{\underline{p}}_{ij}) = \mathbb{E}\,\frac{1}{\underline{N}_i} p_{ij}(1-p_{ij}) \tag{15a}$$

$$\operatorname{cov}(\hat{\underline{p}}_{ij},\hat{\underline{p}}_{i\ell}) = -\mathbb{E}\,\frac{1}{\underline{N}_i} p_{ij}p_{i\ell}. \tag{16a}$$

Bewijs: In tegenstelling tot de bewijsvoering bij lemma 6.1 is het nu noodzakelijk om de struktuur van P te gebruiken. Aangezien $p_{i_1i_2} = 0$ wanneer $i_2 \leq i_1$ hangt $\underline{N}_{ij}$ $(j > i)$ slechts af van $\underline{N}_i$. Daardoor mag ook in deze situatie $\underline{N}_{ij}$ worden opgevat als de stochastische uitkomst van een multinomiaal experiment met parameters $\underline{N}_i$ en $p_{i_1},\ldots,p_{i_2}$, waarbij door de realisatie van $\underline{N}_i$ de grootte N_i van het experiment wordt vastgelegd. Het bewijs verloopt verder analoog als in lemma 6.1. □

In de formules ($13^a,\ldots,16^a$) zijn $\mathbb{E}\,\frac{1}{\underline{N}_i}$, p_{ij} en $p_{i\ell}$ alle onbekend. Op tijdstip 0 echter zijn N_i, N_{ij} en $N_{i\ell}$ $(i,j,\ell \in S)$ bekend doordat de realisaties in het historische proces $\{\underline{N}_i(-\tau) \mid t \in \mathbb{N}\}$ $(i \in S)$ reeds hebben plaatsgehad.
Voor $\operatorname{var}(\hat{\underline{p}}_{ij})$ kan de volgende schatting worden gegeven:

$$s^2(\hat{\underline{p}}_{ij}) := \frac{1}{N_i}\frac{N_{ij}}{N_i}(1 - \frac{N_{ij}}{N_i}) \tag{17}$$

Analoog, voor $\mathrm{cov}(\hat{\underline{p}}_{ij},\hat{\underline{p}}_{i\ell})$:

$$s(\hat{\underline{p}}_{ij},\hat{\underline{p}}_{i\ell}) := -\frac{1}{N_i}\frac{N_{ij}}{N_i}\frac{N_{i\ell}}{N_i} \tag{18}$$

Als schatting voor $\mathbb{E}[(\underline{n}_j(1)-\hat{\underline{n}}_j(1)^2]$ kan met behulp van (6), (10) en (12) de volgende uitdrukking worden gevonden, waarbij p_{ij} wordt geschat door $\frac{N_{ij}}{N_i}$:

$$\sum_{i=1}^{k+1} n_i^2(0).s^2(\hat{\underline{p}}_{ij}) + \sum_{i=1}^{k+1}\frac{n_i(0)}{N_i}(1 - \frac{N_{ij}}{N_i})\frac{N_{ij}}{N_i} =$$

$$= \sum_{i=1}^{k+1}(\frac{n_i^2(0)}{N_i} + n_i(0))\frac{N_{ij}}{N_i}(1 - \frac{N_{ij}}{N_i}) \tag{19}$$

Het quotient van de bijdrage van categorie i tot de verwachte kwadratische fout in de voorspelde bezetting voor categorie j tengevolge van de statistische fout en de bijdrage van categorie i tengevolge van de schattingsfout is volgens formule (19) gelijk aan $\frac{N_i}{n_i(0)}$.
Hieruit volgt, dat de schattingscomponent van de verwachte kwadratische fout kan worden verwaarloosd als b.v. $\frac{N_i}{n_i(0)} > 10$ $(i \in S)$. Dit wil zeggen, dat bij een waargenomen historie van 10 of meer perioden de foutenanalyse kan worden beperkt tot een beschouwing van de statistische fout. In een dergelijke situatie resulteert het verzamelen van historische gegevens over een extra periode in een slechts geringe afname van de kwadratische fout.
We hebben steeds verondersteld dat $\hat{\underline{P}}$ onafhankelijk is van het huidige proces. Is $\hat{\underline{P}}$ een stochastische grootheid die afhangt van het historische proces dan zijn $\hat{\underline{P}}$ en $\underline{n}(0)$ echter i.h.a. afhankelijk.
Voor de situatie waarbij het proces slechts wordt waargenomen gedurende de periode $[-1,0]$, is deze afhankelijkheid als volgt vast te leggen:

$$\underline{n}_j(0) = \sum_{i=1}^{k+1} \underline{N}_{ij} \qquad (j=1,\ldots,k+1)$$

of

$$\sum_{i=1}^{k+1} \underline{N}_i \hat{P}_{ij} = \underline{N}_j(0) \qquad (j=1,\ldots,k+1)$$

Voor de methode is het op zich niet van belang of de startbezetting op tijdstip stochastisch dan wel deterministisch is. Wanneer $\underline{n}(0)$ echter stochastisch ondersteld wordt, is de gegeven afleiding voor $M_E(1)$ moeilijk te generaliseren. Wordt P geschat over een grote historische tijdsperiode dan zal de afhankelijkheid tussen $\underline{n}(0)$ en $\hat{\underline{P}}$ gering zijn, zodat in dat geval (19) een goede benadering is voor de verwachte kwadratische fout. Uit de formules van §6.3 zullen gelijksoortige conclusies volgen voor het geval waarin voorspeld wordt over een grotere periode.

6.2 Voorspellen over meerdere perioden

Wanneer voorspeld wordt over meerdere tijdsperioden geldt:

$$\hat{\underline{n}}(t) = \hat{\underline{n}}(t-1)\hat{\underline{P}} = n(0)\hat{\underline{P}}^t$$

Nu is de verwachte kwadratische fout voor de voorspellingen onder de aannamen van §6.1:

$$M(t) := \mathbb{E}[(\underline{n}(t) - \hat{\underline{n}}(t))' \, (\underline{n}(t) - \hat{\underline{n}}(t))]$$

Ook is nu:

$$M(t) = M_E(t) + M_{ST}(t)$$

waarbij

$$M_E(t) := \mathbb{E}[(\hat{\underline{P}}^t - P^t)' \, n(0)' n(0) (\hat{\underline{P}}^t - P^t)]$$

en

$$M_{ST}(t) := \mathbb{E}[(\underline{n}(t) - \mathbb{E}\,\underline{n}(t))' (\underline{n}(t) - \mathbb{E}\,\underline{n}(t))]$$

Omdat verschillende rijen van $\hat{\underline{P}}^t$ niet meer ongecorreleerd zijn kan de afleiding van §6.1 hier niet worden gevolgd.
Wel geldt vanwege de struktuur van P, dat

$$\mathbb{E}\,\hat{\underline{P}}^t = P^t$$

De historische gegevens over de periode -T, -T+1,...,-1,0 leveren echter, wanneer $t \leq T$, de mogelijkheid om P^t rechtstreeks te schatten.

Laat nu $\underline{N}_{ij}(t)$ het aantal personen zijn, dat gedurende één van de volgende perioden $[-T,-T+t],\ldots,[-t,0]$ overging van categorie i naar categorie j.
Laat verder $\underline{N}_{i}(t) := \sum_{j=1}^{k+1} \underline{N}_{ij}(t)$ zijn, dan geldt voor de maximum likelihoodschatter $\tilde{\underline{P}}_{ij}(t)$ van $p_{ij}(t)$:

$$\tilde{\underline{P}}_{ij}(t) = \frac{\underline{N}_{ij}(t)}{\underline{N}_{i}(t)}$$

De matrix $\tilde{\underline{P}}(t) = \{\tilde{\underline{p}}_{ij}(t)\}$ is een zuivere schatter voor P^t. (NB: $\tilde{\underline{P}}(1)=\hat{\underline{P}}$).
In plaats van met (2^a) kan de bezetting ook worden voorspeld met:

$$\tilde{\underline{n}}(t) = n(0)\ \tilde{\underline{P}}(t) \qquad (20)$$

Bij schatting volgens (20) geldt voor de componenten van M(t):

$$\mathbf{E}\,(\underline{n}_j(t)-\tilde{\underline{n}}_j(t))^2 = \sum_{i=1}^{k+1} n_i(0)\ p_{ij}(t)\ (1-p_{ij}(t)) +$$

$$+ \sum_{i=1}^{k+1} n_i^2(0)\ \mathrm{var}(\tilde{\underline{p}}_{ij}(t))$$

Op analoge wijze als in §6.1 kan bovenstaande uitdrukking geschat worden door:

$$\sum_{i=1}^{k+1} n_i(0)\ \frac{\underline{N}_{ij}(t)}{\underline{N}_i(t)}\left(1 - \frac{\underline{N}_{ij}(t)}{\underline{N}_i(t)}\right) + \sum_{i=1}^{k+1} \frac{n_i^2(0)}{\underline{N}_i(t)}\cdot\frac{\underline{N}_{ij}(t)}{\underline{N}_i(t)}\left(1 - \frac{\underline{N}_{ij}(t)}{\underline{N}_i(t)}\right)$$

Deze formule levert, ook wanneer volgens (2^a) wordt geschat, een goede benadering voor de kwadratische fout, daar $\tilde{\underline{P}}(t)$ en $\hat{\underline{P}}^t$ beide zuivere schatters voor P zijn.
Hierbij dienen natuurlijk wel de slotopmerkingen van §6.1 t.a.v. de afhankelijkheid van $\tilde{\underline{P}}(t)$ en $\underline{n}(0)$ te worden betrokken.

6.3 De Bayesiaanse schattingsmethode

In deze deelparagraaf zal worden aangenomen, dat $\underline{P}$ stochastisch is.
Verondersteld wordt, dat rij i van $\underline{P}$ een apriori verdeling heeft, gegeven door de volgende multivariate beta-kansdichtheid:

$$f(p_{i1},\ldots,p_{i\,k+1}) = \frac{\Gamma\left(\sum_{j=1}^{k+1} b_{ij}\right)}{\prod_{j=1}^{k+1}\Gamma(b_{ij})} \prod_{j=1}^{k+1} p_{ij}^{\,b_{ij}-1} \qquad (22)$$

met $0 \leq p_{ij} \leq 1$ en $\sum_{j=1}^{k+1} p_{ij} = 1 \quad (i \in S)$

Als in een historische waarnemingsperiode N_{ij} overgangen van categorie i naar categorie j plaatsgevonden hebben (i,j $\in$ S), geldt voor de likelihood functie van $p_{i,1},\ldots,p_{i,k+1}$:

$$L\ (p_{i,1},\ldots,p_{i,k+1} \mid N_{i1},\ldots,N_{i,k+1}) \propto \prod_{j=1}^{k+1} p_{ij}^{N_{ij}}$$

Volgens het theorema van Bayes geldt voor de aposteriori kansdichtheid $\pi(p_{i1},\ldots,p_{i,k+1} \mid N_{i1},\ldots,N_{i,k+1})$:

$$\pi(p_{i1},\ldots,p_{i,k+1} \mid N_{i1},\ldots,N_{i,k+1}) \propto f(p_{i1},\ldots,p_{i,k+1}) \cdot (p_{i1},\ldots,p_{i,k+1} \mid N_{i1},\ldots,N_{i,k+1})$$

$$\propto \prod_{j=1}^{k+1} p_{ij}^{b_{ij}+N_{ij}-1}$$

zodat we vinden:

$$\pi(p_{i1},\ldots p_{i,k+1} \mid N_{i1}\ldots N_{i,k+1}) = \frac{\Gamma(\sum_{j=1}^{k+1} b_{ij}+N_{ij})}{\prod_{j=1}^{k+1} \Gamma(b_{ij}+N_{ij})} \prod_{j=1}^{k+1} p_{ij}^{b_{ij}+N_{ij}-1} \qquad (23)$$

De aposteriori-verdeling is weer een multivariate beta-verdeling, welke wordt gekarakteriseerd door:

$$\mathbb{E}\ \underline{p}_{ij} = \frac{b_{ij}+N_{ij}}{\sum_{j=1}^{k+1} b_{ij}+N_{ij}} \qquad (24)$$

$$\text{var}\ \underline{p}_{ij} = \frac{\mathbb{E}\ \underline{p}_{ij}(1-\mathbb{E}\ \underline{p}_{ij})}{\sum_{j=1}^{k+1} (b_{ij}+N_{ij})+1} \qquad (25)$$

$$\text{en}\ \text{cov}(\underline{p}_{ij},\underline{p}_{i\ell}) = \frac{-\ \mathbb{E}\ \underline{p}_{ij}\ \mathbb{E}\ \underline{p}_{i\ell}}{\sum_{j=1}^{k+1} (b_{ij}+N_{ij})+1} \qquad (\ell \neq j)$$

We generaliseren nu de recurrente uitdrukking (5) voor cov $(\underline{n}_i(t+1),\underline{n}_j(t+1))$ tot de situatie waarin P onbekend is. We formuleren daartoe het volgende lemma.

Lemma 6.3:

Wanneer een Markov-proces beschouwd wordt met daarbij gegeven de oneigenlijke multivariate-beta-apriori-verdeling ($b_{ij}=0$, $i,j = 1,\ldots,k+1$) voor $\underline{P}$ en wanneer gegevens over de bezetting N_i in categorie i en personeelsstromen N_{ij} beschikbaar zijn, dan geldt:

$$\begin{aligned}\operatorname{cov}(\underline{n}_i(t+1),\underline{n}_j(t+1)) = &\sum_{\ell=1}^{k+1}\sum_{r=1}^{k+1} \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{rj}\ \operatorname{cov}(\underline{n}_\ell(t),\underline{n}_r(t)) \\ &+ \sum_{\ell=1}^{k+1} (\delta_{ij}-\mathbb{E}\ \underline{p}_{\ell j})\ \mathbb{E}\,\underline{p}_{\ell i}\ \left(\frac{N_\ell+\mathbb{E}\,\underline{n}_\ell(t)}{N_\ell+1}\right)\ \mathbb{E}\ \underline{n}_\ell(t) \\ &+ \sum_{\ell=1}^{k+1} (\delta_{ij}-\mathbb{E}\ \underline{p}_{\ell j})\ \mathbb{E}\ \underline{p}_{\ell i}\ \frac{\operatorname{var}(\underline{n}_\ell(t))}{N_\ell+1} \qquad (26)\end{aligned}$$

Bewijs:

A posteriori geldt volgens het Bayesiaanse standpunt, dat $\underline{p}_{ij}$ een multivariate beta-kansdichtheid heeft met verwachting $\mathbb{E}\ \underline{p}_{ij}$.
Daaruit volgt:

$$\text{i)}\quad \mathbb{E}\ [\underline{n}_{ij}(t) \mid \underline{n}_i(t) = n_i(t)] = n_i(t)\ \mathbb{E}\ \underline{p}_{ij}(t)$$

$$\text{ii)}\quad \mathbb{E}\ [\underline{n}_{\ell i}(t)\ \underline{n}_{rj}(t) \mid \underline{n}_\ell(t) = n_\ell(t),\ \underline{n}_r(t) = n_r(t)] =$$

$$= n_\ell(t)\ n_r(t)\ \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{rj} \qquad (\ell \neq r)$$

Impliciet is hier aangenomen dat $\underline{p}_{\ell i}$ en $\underline{p}_{rj}$ $(\ell \neq r)$ onafhankelijk zijn.
Tevens geldt (zie [2] en [6] voor het analogon in de situatie waar $\underline{P} = P$):

$$\text{iii)}\quad \mathbb{E}\ [\underline{n}_{\ell i}(t)\underline{n}_{\ell j}(t) \mid \underline{n}_\ell(t) = n_\ell(t)] =$$

$$= n_\ell^2(t)\ \mathbb{E}\ \underline{p}_{\ell i}\ \underline{p}_{\ell j} + n_\ell(t)\ (\delta_{ij}\ \mathbb{E}\ \underline{p}_{\ell i} - \mathbb{E}\ \underline{p}_{\ell i}\underline{p}_{\ell j})$$

Hieruit volgt voor de onvoorwaardelijke verwachtingen

$$\text{i')}\quad \mathbb{E}\ \underline{n}_{ij}(t) = \mathbb{E}\ \underline{n}_i(t)\ \mathbb{E}\ \underline{p}_{ij}.$$

ii') $\mathbb{E}\ \underline{n}_{\ell i}(t)\ \underline{n}_{rj}(t) = \mathbb{E}\ \underline{n}_{\ell}(t)\ \underline{n}_{r}(t)\ \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{rj} \quad (\ell \neq r)$

iii') $\mathbb{E}\ \underline{n}_{\ell i}(t)\ \underline{n}_{\ell j}(t) = \mathbb{E}\ \underline{n}_{\ell}^{2}(t)\ \mathbb{E}\ \underline{p}_{\ell i}\underline{p}_{\ell j} +$

$+ \mathbb{E}\ \underline{n}_{\ell}(t)\ (\delta_{ij}\ \mathbb{E}\ \underline{p}_{\ell i} - \mathbb{E}\ \underline{p}_{\ell i}\underline{p}_{\ell j})$

Vanwege de relatie $\underline{n}_{i}(t+1) = \sum_{\ell=1}^{k+1} \underline{n}_{\ell i}(t)$ volgt nu:

$$\operatorname{cov}(\underline{n}_{i}(t+1), \underline{n}_{j}(t+1)) = \mathbb{E}\ \underline{n}_{i}(t+1)\underline{n}_{j}(t+1) - \mathbb{E}\ \underline{n}_{i}(t+1)\ \mathbb{E}\,\underline{n}_{j}(t+1)$$

$$= \sum_{\ell=1}^{k+1}\sum_{r=1}^{k+1} [\mathbb{E}\ \underline{n}_{\ell i}(t)\underline{n}_{rj}(t) - \mathbb{E}\ \underline{n}_{\ell i}(t)\ \mathbb{E}\ \underline{n}_{rj}(t)]$$

$$= \sum_{\ell=1}^{k+1}\sum_{r=1}^{k+1} (\mathbb{E}\ \underline{n}_{\ell}(t)\ \underline{n}_{r}(t) - \mathbb{E}\ \underline{n}_{\ell}(t)\ \mathbb{E}\ \underline{n}_{r}(t))\ \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\,\underline{p}_{rj}$$

$$+ \sum_{\ell=1}^{k+1} \mathbb{E}\ \underline{n}_{\ell}^{2}(t)\ \operatorname{cov}(\underline{p}_{\ell i}\ \underline{p}_{\ell j}) +$$

$$+ \sum_{\ell=1}^{k+1} \mathbb{E}\ n_{\ell}(t)\ (\delta_{ij}\ \mathbb{E}\ \underline{p}_{\ell i} - \mathbb{E}\ \underline{p}_{\ell i}\underline{p}_{\ell j})$$

$$= \sum_{\ell=1}^{k+1} \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{rj}\ \operatorname{cov}(\underline{n}_{\ell}(t)\underline{n}_{r}(t)) +$$

$$+ \sum_{\ell=1}^{k+1} (\delta_{ij}\ \mathbb{E}\ \underline{p}_{\ell j} - \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{\ell j})\ \frac{\operatorname{var}(\underline{n}_{\ell}(t))}{N_{\ell}+1} +$$

$$+ \sum_{\ell=1}^{k+1} (\delta_{ij}\ \mathbb{E}\ \underline{p}_{\ell j} - \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{\ell j})\ \frac{(\mathbb{E}\ \underline{n}_{\ell}(t))^{2}}{N_{\ell}+1}$$

$$+ \sum_{\ell=1}^{k+1} \mathbb{E}\ \underline{n}_{\ell}(t)\ (\delta_{ij}\ \mathbb{E}\ \underline{p}_{\ell i} - \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{\ell j} - \operatorname{cov}\ \underline{p}_{\ell i}, \underline{p}_{\ell j}))$$

$$= \sum_{\ell=1}^{k+1}\sum_{r=1}^{k+1} \mathbb{E}\ \underline{p}_{\ell i}\ \mathbb{E}\ \underline{p}_{rj}\ \operatorname{cov}\ (\underline{n}_{\ell}(t)\underline{n}_{r}(t)) +$$

$$+ \sum_{\ell=1}^{k+1} (\delta_{ij} - \mathbb{E}\ \underline{p}_{\ell i})\ \mathbb{E}\ \underline{p}_{\ell j}\ \frac{\operatorname{var}(\underline{n}_{\ell}(t))}{N_{\ell}+1}$$

$$+ \sum_{\ell=1}^{k+1} (\delta_{ij} - \mathbb{E}\ \underline{p}_{\ell i})\ \mathbb{E}\ \underline{p}_{\ell j}\ \frac{(\mathbb{E}\ \underline{n}_{\ell}(t))^{2}}{N_{\ell}+1}$$

$$+ \sum_{\ell=1}^{k+1} \frac{N_{\ell+1}^{-1}}{N_\ell} \; \mathbb{E}\, \underline{n}_\ell(t) \; (\delta_{ij} \; \mathbb{E}\, \underline{p}_{\ell i} - \mathbb{E}\, \underline{p}_{\ell i} \, \underline{p}_{rs}) \qquad \square$$

Beschouwen we de situatie voor t = 1 nader. Wanneer we in (24) $\frac{N_{ij}}{N_i}$ voor $\mathbb{E}\, \hat{\underline{p}}_{ij}$ substitueren geldt:

$$\text{var}\,(\underline{n}_j(1)) = \sum_{\ell=1}^{k+1} (1 - \frac{N_{\ell j}}{N_\ell}) \frac{N_{\ell j}}{N_\ell} \frac{N_\ell + n_\ell(0)}{N_\ell + 1} n_\ell(0) \qquad (27)$$

De introductie van de stochastiek in het model heeft tot gevolg dat var $(\underline{n}_j(1))$ met een factor $\frac{N_{\ell+} n_\ell(0)}{N_\ell + 1}$ wordt vermenigvuldigd.

Bij de frequentistische aanpak werd als vermenigvuldigingsfactor $(1 + \frac{n_\ell(0)}{N_\ell})$ gevonden.

Daar de orde grootte van deze factoren dezelfde is, leveren beide methoden dezelfde conclusies. Voor de frequentistische methode werden in §6.1 deze conclusies reeds vermeld.

Beide methoden hebben als nadeel, dat generalisatie van (19) resp. (27) tot het geval t > 1 heel bewerkelijk is. Bij de Bayesiaanse methode is de reden hiervan, dat hogere-orde momenten van $\underline{P}$ een rol gaan spelen in de uitdrukkingen.

7. Statistische aspecten van het aggregeren van categorieën

7.1. Inleiding

Laat $W \subset S$ een verzameling categorieën zijn. Aangenomen wordt nu, dat voor een willekeurige categorie $i \in W$ de overgangskans naar een willekeurige vaste categorie $j \in S\backslash W$ weinig varieert met de keuze van $i \in W$.
Tevens wordt aangenomen, dat de categorieën $i \in W$ een geringe bezetting hebben.
In deze paragraaf wordt bestudeerd, wat het effect is van het samenvoegen van alle categorieën in W tot één enkele categorie terwijl daarbij ook de historische gegevens over de overgangen geaggregeerd worden.
In §3 is reeds nagegaan wat het effect is van een variatie van bepaalde overgangspercentages. Op de verwachte bezetting, wanneer de overgangspercentages bekend zijn. De daar geformuleerde resultaten geven bovengrenzen in norm voor de voorspelfout bij variatie van overgangspercentages.
In deze paragraaf wordt voor alle overgangen vanuit categorieën $i \in W$ naar een categorie $j \in S\backslash W$ één overgangspercentage berekend, gebaseerd op de geaggregeerde historische gegevens. In §7.2 wordt de verwachte kwadratische fout in de voorspellingen op tijdstip 1 beschouwd in deze situatie. Voorts wordt in §7.3 nagegaan onder welke voorwaarden de fout die door te aggregeren wordt geïntroduceerd, van dezelfde grootte orde is als de schattingsfout. De statistische benaderingswijze in deze paragraaf is gebaseerd op Bayesiaanse methoden.

7.2. De verwachte kwadratische fout

Voor categorieën $i \in S\backslash W$ geldt, dat de aposteriori kansdichtheid van de vector $(\underline{p}_{i1},\underline{p}_{i2},\ldots,\underline{p}_{i,k+1})$ (zie (22) met $b_{ij} = 0$, $i,j \in S$) gelijk is aan:

$$\pi(p_{i1},\ldots,p_{i,k+1}) = \frac{\Gamma(\sum_{j=1}^{k+1} N_{ij})}{\prod_{j=1}^{k+1} \Gamma(N_{ij})} \prod_{j=1}^{k+1} p_{ij}^{N_{ij}-1}$$

Aangenomen wordt, dat voor het aggregaat van categorieën W overgangswaarschijnlijkheden $\underline{p}^*_{w1},\ldots,\underline{p}^*_{w,k+1}$ bestaan, met a posteriori verdeling:

$$\pi(p^*_{w1},\ldots,p^*_{w,k+1}) = \frac{\Gamma(\sum_{i\in W}\sum_{j=1}^{k+1} N_{ij})}{\prod_{j=1}^{k+1} \Gamma(\sum_{i\in W} N_{ij})} \prod_{j=1}^{k+1} p^{*(\sum_{i\in W} N_{ij})-1}_{wj}$$

Worden N_{wj} en N_w als volgt gedefinieerd

$$N_{wj} := \sum_{i \in W} N_{ij}$$

$$N_w := \sum_{i \in W} \sum_{j=1}^{k+1} N_{ij} = \sum_{j-1}^{k+1} N_{wj}$$

dan geldt:

$$\pi(p^*_{w1},\ldots,p^*_{w,k+1}) = \frac{\Gamma(N_w)}{\prod_{j=1}^{k+1} \Gamma(N_{wj})} \prod_{j=1}^{k+1} p^{*\,N_{wj}-1}_{wj} \qquad (28)$$

Voor de aposteriori verwachting en variantie volgt hieruit:

$$\mathbb{E}\ \underline{p}^*_{wj} = \frac{N_{wj}}{N_w} \qquad j \notin W$$

$$0 \qquad j \in W$$

$$\operatorname{var} \underline{p}^*_{wj} = \frac{\mathbb{E}\ \underline{p}^*_{wj}(1 - \mathbb{E}\ \underline{p}^*_{wj})}{N_w+1}$$

$$\operatorname{cov}(\underline{p}^*_{wj}, \underline{p}^*_{w\ell}) = \frac{-\ \mathbb{E}\ \underline{p}^*_{wj}\ \mathbb{E}\ \underline{p}^*_{w\ell}}{N_w+1} \qquad (\ell \neq j)$$

Voor de verwachte kwadratische fout op tijdstip 1 in positie j volgt, wanneer $\underline{N}_{wj}$ het aantal overgangen vanuit W wordt genoemd bij apriori verdeling (28) voor $p^*_{w1},\ldots,p^*_{w,k+1}$, de volgende uitdrukking.

$$\mathbb{E}\Big[\sum_{\ell \notin W} (\underline{n}_{\ell j}(0) - n_\ell(0)\ \mathbb{E}\ \underline{p}_{\ell j}) + \underline{n}_{wj}(0) - n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj}\Big]^2 +$$

$$+ \Big(n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj} - \sum_{i \in W} n_i(0)\ \mathbb{E}\ \underline{p}_{ij}\Big)^2 \qquad (j \in W)$$

$$= \sum_{\ell \notin W} (1 - \mathbb{E}\ \underline{p}_{\ell j})\ \mathbb{E}\ \underline{p}_{\ell j} \left(\frac{N_\ell + N_\ell(0)}{N_{\ell+1}}\right) n_\ell(0) +$$

$$+ (1 - \mathbb{E}\ \underline{p}^*_{wj})\ \mathbb{E}\ \underline{p}^*_{wj} \left(\frac{N_w + N_w(0)}{N_{w+1}}\right) n_w(0) +$$

$$+ (n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj} - \sum_{i\in W} n_i(0)\ \mathbb{E}\ \underline{p}_{ij})^2 \qquad (j \notin W) \qquad (29)$$

De term $(n_w(0)\ \mathbb{E}\ \underline{p}_{wj} - \sum_{i\in W} n_i(0)\ \mathbb{E}\ \underline{p}_{wj})^2$ is hier de kwadratische systematische fout, geïntroduceerd vanwege het feit, dat $\underline{n}_{wj}(0)$ en $\sum_{i\in W} \underline{n}_{ij}(0)$ niet dezelfde verwachtingswaarde hebben.
Wanneer $\mathbb{E}\ \underline{p}_{ij} = \mathbb{E}\ \underline{p}^*_{wj}$ $(i \in W,\ j \notin W)$ vervalt deze systematische fout.

7.3. De grootte van de agregatiefout

De verwachte kwadratische fout A_j in de voorspelde bezetting van categorie j op tijdstip 1, wanneer de categorieën binnen $W \in S$ worden geaggregeerd, wordt in deze deelparagraaf vergeleken met de verwachte kwadratische fout B_j in de voorspelde bezetting van categorie j op tijdstip 1 wanneer de categorieën binnen W niet geaggregeerd worden.
Volgens (29) geldt:

$$A_j = \mathbb{E}\ [\sum_{\ell=1,\ell\in W}^{k+1} (\underline{n}_{\ell j}(0) - n_\ell(0)\ \mathbb{E}\ \underline{p}_{\ell j}) + \underline{n}_{wj}(0) - n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj}]^2 +$$

$$+ (n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj} - \sum_{i\in W} n_i(0)\ \mathbb{E}\ \underline{p}_{ij})^2$$

$$= \sum_{\ell=1,\ell\notin W}^{k+1} \mathbb{E}\,[\underline{n}_{\ell j}(0) - n_\ell(0)\ \mathbb{E}\ \underline{p}_{\ell j}]^2 +$$

$$= \mathbb{E}\,[\underline{n}_{wj}(0) - n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj}]^2 +$$

$$+ (n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj} - \sum_{i\in W} n_i(0)\ \mathbb{E}\ \underline{p}_{ij})^2 \qquad (j \notin W)$$

en voor B_j volgt nu:

$$B_j = \mathbb{E}\,[\sum_{\ell=1}^{k+1} \underline{n}_{\ell j}(0) - n_\ell(0)\ \mathbb{E}\ \underline{p}_{\ell j}]^2$$

$$+ \sum_{\ell=1}^{k+1} \mathbb{E}\,[\underline{n}_{\ell j}(0) - n_\ell(0)\ \mathbb{E}\ \underline{p}_{\ell j}]^2 \qquad (j \notin W)$$

Het verschil van A_j en B_j is gelijk aan:

$$A_j - B_j = \sum_{\ell \in W} \mathbb{E}\ (\underline{n}_{\ell j}(0) - n_\ell(0)\ \mathbb{E}\ \underline{p}_{\ell j})^2 +$$

$$- \mathbb{E}\ (\underline{n}_{wj}(0) - n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj})^2 +$$

$$- (n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj} - \sum_{i \in W} n_i(0)\ \mathbb{E}\ \underline{p}^*_{ij})^2$$

$$= \sum_{\ell \in W} n_\ell(0)\ \frac{N_\ell + N_\ell(0)}{N_\ell + 1}\ \mathbb{E}\ \underline{p}_{\ell j}\ (1 - \mathbb{E}\ \underline{p}_{\ell j}) +$$

$$- n_w(0)\ \frac{N_w + N_w(0)}{N_w + 1}\ \mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj})$$

$$- (n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj} - \sum_{i \in W} n_i(0)\ \mathbb{E}\ \underline{p}_{ij})^2. \qquad (j \notin W)$$

Opgemerkt kan worden dat $A_j - B_j$ slechts afhangt van de verwachte kwadratische fout, gemaakt bij de voorspelling van personeelsstromen van categorieën $\ell \in W$ naar categorie j. $(j \notin W)$.
Eigenlijk valt $A_j - B_j$ in drie termen uiteen, t.w.

$$A_j - B_j = C_{1j} - C_{2j} - C_{3j}$$

met

$$C_{ij} := \sum_{\ell \in W} n_\ell(0)\ \frac{N_\ell + n_\ell(0)}{N_\ell + 1}\ \mathbb{E}\ \underline{p}_{\ell j}\ (1 - \mathbb{E}\ \underline{p}_{\ell j})$$

de verwachte kwadratische fout tengevolge van de overgangen van de verschillende categorieën uit W naar categorie j, als geen aggregatie plaats heeft, verder met

$$C_{2j} := n_w(0)\ \frac{N_w + n_w(0)}{N_w + 1}\ \mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj})$$

de verwachte kwadratische fout tengevolge van overgangen vanuit W naar categorie j, wanneer wordt geaggregeerd.
tenslotte met

$$\underline{p}_{wj} \qquad i \in W$$

de systematische kwadratische fout, die bij voorspelling wordt gemaakt als aggregatie van categorieën plaatsheeft.
In het volgende lemma wordt een voorwaarde gegeven, waaronder geldt, dat

$$C_{3j} < \frac{n_w(0)}{N_w+n_w(0)}\, C_{2j} \qquad (j \notin W)$$

Lemma 7.1.

Onder de voorwaarde, dat:

$$\left(\frac{n_\ell(0)}{n_w(0)} - \frac{N_\ell}{N_w}\right)^2 (\mathbb{E}\ \underline{p}_{\ell j} - \mathbb{E}\ \underline{p}^*_{wj})^2 <$$

$$\frac{\mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ p^*_{wj})}{(\#W)^2\ (N_w+1)} \qquad (\ell \in W, j \notin W)$$

geldt:

$$C_{3j} < \frac{n^2{}_w(0)}{N_w+1}\ \mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj}) \qquad (j \notin W)$$

bewijs: Daar $\mathbb{E}\ \underline{p}^*_{wj} = \sum_{i\in W} \frac{N_i}{N_w}\ \mathbb{E}\ \underline{p}_{ij}$ volgt, dat:

$$C_{3j} = [\sum_{i\in W} n_i(0)\ \mathbb{E}\ \underline{p}_{ij} - n_w\ \mathbb{E}\ \underline{p}^*_{wj}]^2 =$$

$$= [\sum_{i\in W} n_i(0)\ (\mathbb{E}\ \underline{p}_{ij} - \mathbb{E}\ \underline{p}^*_{wj})]^2 =$$

$$= n_w^2(0)\ [\sum_{i\in W} \frac{n_i(0)}{n_w(0)}\ (\mathbb{E}\ \underline{p}^*_{wj} - \mathbb{E}\ \underline{p}_{ij})]^2 =$$

$$= nw^2(0)\ [\sum_{i\in W} \left(\frac{n_i(0)}{n_w(0)} - \frac{N_i}{N_w}\right)\ (\mathbb{E}\ \underline{p}^*_{wj} - \mathbb{E}\ \underline{p}_{ij})]^2 =$$

$$< nw^2(w) \sum_{i\in W} \left[\frac{[\mathbb{E}\ \underline{p}^*_{wj}(1 - \mathbb{E}\ p_{wj})]^{\frac{1}{2}}}{\#W \cdot (N_w+1)^{\frac{1}{2}}}\right]^2$$

$$= \frac{nw^2(0)}{N_w+1}\ \mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj}). \qquad \square$$

Verder geldt voor $\mathbb{E}\ \underline{p}^*_{wj}(1 - \mathbb{E}\ \underline{p}^*_{wj})$ de volgende eigenschap:

Lemma 7.3:

$$\mathbb{E}\ \underline{p}^*_{wj}(1 - \mathbb{E}\ \underline{p}^*_{wj}) \geq \sum_{i \in W} \frac{N_i}{N_w}\, \mathbb{E}\ \underline{p}_{ij}(1 - \mathbb{E}\ \underline{p}_{ij})$$

Bewijs:

Omdat $\mathbb{E}\ \underline{p}^*_{wj} = \sum_{i \in W} \frac{N_i}{N_w}\, \mathbb{E}\ \underline{p}_{ij}$ volgt

$$\mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj}) = \mathbb{E}\ \underline{p}^*_{wj} - (\mathbb{E}\ \underline{p}^*_{wj})^2 =$$

$$= \sum_{i \in W} \frac{N_i}{N_w}\, \mathbb{E}\ \underline{p}_{ij} - \Big(\sum_{i \in W} \frac{N_i}{N_w}\, \mathbb{E}\ \underline{p}_{ij}\Big)^2.$$

Nu is op grond van de convexiteitseigenschap:

$$(\mathbb{E}\ \underline{p}^*_{wj})^2 = \Big(\sum_{i \in W} \frac{N_i}{N_w}\, \mathbb{E}\ \underline{p}_{ij}\Big)^2$$

$$\leq \sum_{i \in W} \frac{N_i}{N_w} \cdot (\mathbb{E}\ \underline{p}_{ij})^2$$

zodat

$$\mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj}) \geq \sum_{i \in W} \frac{N_i}{N_w}\, \mathbb{E}\ \underline{p}_{ij} - (\mathbb{E}\ \underline{p}_{ij})^2$$

$$= \sum_{i \in W} \frac{N_i}{N_w}\, \mathbb{E}\ \underline{p}_{ij}\ (1 - \mathbb{E}\ \underline{p}_{ij}) \qquad \square$$

Als voldaan is aan de volgende condities:

$$\Big|\frac{2n_w(0)-1}{N_w+1}\Big| < \beta,$$

$$\Big|\frac{n_\ell(0)-1}{n_\ell+1}\Big| < \beta$$

voor zekere $0 < \beta < 1$, met β voldoende klein, dan mag de fountenanalyse worden beperkt tot de statistische fout. Er geldt dan bij benadering:

$$|A_j - B_j| \approx \Big| \sum_{\ell \in W} n_\ell(0)\ \mathbb{E}\, \underline{p}_{\ell j}\ (1 - \mathbb{E}\ \underline{p}_{\ell j}) +$$

$$+ n_w(0)\ \mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj})\Big|$$

$$= \Big| \sum_{\ell \in W} n_\ell(0) [\mathbb{E}\ \underline{p}_{\ell j}(1 - \mathbb{E}\ \underline{p}_{\ell j}) - \mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj})] \Big|$$

$$= n_w(0) \Big| \sum_{\ell \in W} \frac{n_\ell(0)}{n_w(0)}\ \mathbb{E}\ \underline{p}_{\ell j}\ (1 - \mathbb{E}\ \underline{p}_{\ell j}) - \mathbb{E}\ \underline{p}^*_{wj}\ (1 - \mathbb{E}\ \underline{p}^*_{wj}) \Big|$$

$$\leq n_w(0) \sum_{\ell \in W} \frac{n_\ell(0)}{n_w(0)} \Big| \mathbb{E}\ \underline{p}_{\ell j} - \mathbb{E}\ \underline{p}^*_{wj} \Big| .$$

Als nu voor zekere $\varepsilon > 0$:

$$\text{i)}\quad \frac{n_\ell(0)}{n_w(0)} \Big| \mathbb{E}\ \underline{p}_{\ell j} - \mathbb{E}\ \underline{p}^*_{wj} \Big| < \varepsilon \cdot \frac{1}{\#W}\ \mathbb{E}\ \underline{p}^*_{wj} \cdot\ (1 - \mathbb{E}\ \underline{p}^*_{wj})$$

$$\text{ii)}\quad \Big| \mathbb{E}\ \underline{p}_{\ell j} - \mathbb{E}\ \underline{p}_{wj} \Big| < \varepsilon \cdot \mathbb{E}\ \underline{p}_{\ell j}\ (1 - \mathbb{E}\ \underline{p}_{\ell j})$$

geldt voor de aggregatiefout:

$$|A_j - B_j| \approx \varepsilon \cdot C_{2j} \qquad (j \notin W)$$

$$|A_j - B_j| \approx \varepsilon \cdot C_{ij} \qquad (j \notin W)$$

Samenvattend kan worden geconcludeerd dat als:

$$\text{i)}\quad \Big(\frac{n_\ell(0)}{n_w(0)} - \frac{N_\ell}{N_w}\Big)^2\ (\mathbb{E}\ \underline{p}_{\ell j} - \mathbb{E}\ \underline{p}_{wj})^2 < \frac{\mathbb{E}\ \underline{p}_{wj}\ (1 - \mathbb{E}\ \underline{p}_{wj})}{(\#\ W)^2 . N_w + 1}$$

de systematische kwadratische fout kan worden verwaarloosd t.o.v. de bijdrage van de verwachte kwadratische fout in de voorspelde bezetting van categorie j op tijdstip 1 tengevolge van overgangen vanuit W naar j, wanneer aggregatie plaatsvindt.

Als bovendien voor zekere $0 < \beta << 1$:

$$\text{ii)}\quad \Big| \frac{2n_w(0) - 1}{N_w + 1} \Big| < \beta$$

$$\text{en}\quad \Big| \frac{n_\ell(0) - 1}{N_\ell + 1} \Big| < \beta$$

kan ook de agrregatiefout worden verwaarloosd t.o.v. genoemde bijdrage van de verwachte kwadratische fout.

Of aggregatie, wanneer aan i) en ii) is voldaan, de verwachte kwadratische fout vergroot, hangt verder af van de minimale waarde van $\varepsilon > 0$ waarvoor voldaan is aan:

iii) $n_\ell(0)$ $\left| \mathbb{E}\ P_{\ell j} - \mathbb{E}\ P^*_{wj} \right| < \varepsilon \cdot \frac{1}{\# W} \cdot \mathbb{E}\ P^*_{wj} \left(1 - \mathbb{E}\ P^*_{wj} \right)$

en $\left| \mathbb{E}\ P_j - \mathbb{E}\ P^*_{wj} \right| < \varepsilon \cdot \mathbb{E}\ P_{\ell j} \left(1 - \mathbb{E}\ P_{\ell j} \right)$

Literatuur

[1] Anderson, T.W. en Goodman L.A. (1954)
Statistical Inference About Markov Chains.The Annals of Mathematical statistics 28 p. 89-110.

[2] Bartholomew, D.J. (1973)
Stochastic Models for Social Processes (2nd ed). John Wiley and Sons, New York.

[3] Bartholomew, D.J. (1975)
Errors of Prediction for Markov Chain Models. Journal of the Royal Statistical Society (B) 37 p. 444-456.

[4] Lee, T.C., Judge G.G., Zellner A. (1970)
Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data. North Holland Publishing Company, Amsterdam.

[5] Wessels, J., van Nunen, J.A.E.E. (1976)
FORMASY, Forecasting and Recruitment in Manpower Systems. Statistica Neerlandica 30 p. 173-193.

[6] Esser, F.L.G. (1975)
Voorspellen en recruteren in personeelssystemen.Afstudeerverslag, Technische Hogeschool Eindhoven.