A componentwise PageRank algorithm

Christopher Engström and Sergei Silvestrov

Division of Applied Mathematics
The School of Education, Culture and Communication (UKK)
Mälardalen University, Box 883, 721 23 Västerås, Sweden
(E-mail: christopher.engstrom@mdh.se, sergei.silvestrov@mdh.se)

Abstract. In this article we will take a look at a variant of the PageRank algorithm initially used by S. Brinn and L. Page to rank homepages on the Internet. The aim of the article is to see how we can use the topological structure of the graph to speed up calculations of PageRank without doing any additional approximations. We will see that by considering a non-normalized version of PageRank it is easy to see how we can handle different types of vertices or strongly connected components in the graph more efficiently. Using this we propose two PageRank algorithms, one similar to the Lumping algorithm proposed by Qing et al which handles certain types of vertices faster and last another PageRank algorithm which can handle more types of vertices as well as strongly connected components more effectively. In the last sections we will look at some specific types of components as well as verifying the time complexity of the algorithm.
Keywords: PageRank, strongly connected component, random walk .

1 Introduction

While the PageRank algorithm initially used by S. Brinn and L. Page to rank homepages on the Internet is very efficient [2], networks such as the graph describing the homepages and their links on the Internet are often huge, and further more are quickly growing. This calls for increasingly higher requirements on algorithms working on this kind of data. Some studies have been made in looking at how certain parameters influence the convergence speed and stability, such as how the constant c affect the condition number and convergence speed [6,8]. While many steps have been made to improve the method by for example aggregating webpages that are close [7] or not compute PageRank for pages that are deemed to have already converged [10]. Another method to speed up the algorithm is to remove so called dangling pages (pages with no links to any other page), and then calculate their rank at the end [1,9]. A similar method can also be used for root nodes (pages with no links from any other pages) [12]. The first method is similar to the one proposed by Qing Yu et al [12], while the second aims to improve the method further. Specifically we will look at the topological structure of the graph and see which types of vertices or strongly connected components in the graph can be handled more efficiently. To help with this we will need to define a slightly different version of

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

PageRank without the normalization in the original formulation, which allows for us to compare the rank between two different graphs more easily.

While we above talked about webpages and links, in the rest of the article we will mainly consider any simple directed graph G with vertices V (pages) and edges E (links). It is also worth to note that while we assume the basic underlying algorithm used to calculate PageRank is the Power method as described in [3], any method could actually be used. We are merely looking at what part of the graph we need to use such a method for, and for which parts we can do something better. Traditionally PageRank is defined as in Definition. 1.

Definition 1. PageRank $\boldsymbol{R}^{(1)}$ for vertices in system S consisting of n vertices is defined as the (right) eigenvector with eigenvalue one to the matrix:

$$M = c(\mathbf{A} + \boldsymbol{g}\boldsymbol{u}^\top)^\top + (1-c)\boldsymbol{u}\boldsymbol{e}^\top \tag{1}$$

where $\mathbf{A}$ is the adjacency matrix weighted such that the sum over every non-zero row is equal to one (size $n \times n$), $\boldsymbol{g}$ is a $n \times 1$ vector with zeros for vertices with outgoing edges and 1 for all vertices with no outgoing edges, $\boldsymbol{u}$ is a $n \times 1$ non-negative vector with $||\boldsymbol{u}||_1 = 1$, $\boldsymbol{e}$ is a one-vector with size $n \times 1$ and $0 < c < 1$ is a scalar.

In this article we will use a slightly different version of PageRank with the main difference in that it does not normalize the rank, resulting in an easier handling of multiple graphs and their rank.

Definition 2. ([4]) PageRank $\boldsymbol{R}^{(3)}$ for system S is defined as

$$\boldsymbol{R}^{(3)} = \frac{\boldsymbol{R}^{(1)}||\boldsymbol{v}||_1}{d},\ d = 1 - \sum c\mathbf{A}^\top \boldsymbol{R}^{(1)} \tag{2}$$

where $\boldsymbol{v}$ is a non-negative weight vector such that $\boldsymbol{u} \propto \boldsymbol{v}$.

The main difference between $\boldsymbol{R}^{(1)}$ and $\boldsymbol{R}^{(3)}$ is that the rank between two different systems can freely be compared between them in $\boldsymbol{R}^{(3)}$ while this is not easily done in $\boldsymbol{R}^{(1)}$ even if the two graphs have the same number of vertices, this because the rank is normalized depending on number of vertices in a system . Note that while $\boldsymbol{v}$ and $\boldsymbol{u}$ in Definition. 1 need to be proportional ($\boldsymbol{v} \propto \boldsymbol{v}$), $\boldsymbol{v}$ can be scaled in any way deeped appropriately as long as it is still a non-negative. Specifically this means that we do no need to re-scale $\boldsymbol{v}$ when adding or subtracting a vertex from the system. Similarly we note that the only difference between $\boldsymbol{R}^{(1)}$ and $\boldsymbol{R}^{(3)}$ is in the scaling, thus the two versions are obviously proportional to eachother. ($\boldsymbol{R}^{(3)}$ is written with a "3" rather than "2"in order to keep it consistent with our other work).

$\boldsymbol{R}^{(1)}$ can be defined as the stationary distribution of some Markov chain, similarly $\boldsymbol{R}^{(3)}$ can also be defined by considering a random walk on a graph.

Definition 3. Consider a random walk on a graph described by $\mathbf{A}$. We walk to a new vertex from our current with probability $0 < c < 1$ and stop the

random walk with probability $1-c$. Then PageRank $\boldsymbol{R}^{(3)}$ for a single vertex can be written as

$$\boldsymbol{R}_j^{(3)} = \left(v_j + \sum_{e_i \in S, e_i \neq e_j} v_i P(e_i \to e_j) \right) \left(\sum_{k=0}^{\infty} (P(e_j \to e_j))^k \right) \tag{3}$$

where $P(e_i \to e_j)$ is the probability to hit node e_j in a random walk starting in node e_i. This can be seen as the expected number of visits to e_j if we do multiple random walks, starting in every vertex a number of times desribed by $\boldsymbol{v}$.

It is easy to prove that $\boldsymbol{R}^{(3)}$ as defined in Definition. 3 and Definition. 2 are equivalent, the proof follows very closely that of a slightly different definition of PageRank in [5].

Theorem 1. *The two definitions Definition. 2 with PageRank $\boldsymbol{R}^{(3)}$ as a scaled version of $\boldsymbol{R}^{(1)}$ and Definition. 3 with PageRank constructed by considering a random walk are equivalent.*

Proof. Starting with Definition. 3:

$$\boldsymbol{R}_j^{(3)} = \left(\sum_{e_i \in S, e_i \neq e_j} v_i P(e_i \to e_j) + v_j \right) \left(\sum_{k=0}^{\infty} (P(e_j \to e_j))^k \right) \quad . \tag{4}$$

$(c\mathbf{A}^\top)^k_{ij}$ is the probability to be in vertex e_i starting in vertex e_j after k steps. Multiplying with the vector $\boldsymbol{v}$ therefor gives the sum of all the probabilities to be in node e_i after k steps starting in every node once weighted by $\boldsymbol{v}$. The expected total number of visits is the sum of all probabilities to be in node e_i for every step starting in every node:

$$\boldsymbol{R}_j^{(3)} = \left(\left(\sum_{k=0}^{\infty} (c\mathbf{A}^\top)^k \right) \boldsymbol{v} \right)_j \quad . \tag{5}$$

$\sum_{k=0}^{\infty} (c\mathbf{A}^\top)^k$ is the Neumann series of $(\mathbf{I} - c\mathbf{A}^\top)^{-1}$ which is guaranteed to converge since $c\mathbf{A}^\top$ is non-negative and have column sum < 1. This gives:

$$\boldsymbol{R}^{(3)} = \left(\left(\sum_{k=0}^{\infty} (c\mathbf{A}^\top)^k \right) \boldsymbol{v} \right) = (\mathbf{I} - c\mathbf{A}^\top)^{-1} \boldsymbol{v} \quad . \tag{6}$$

We continue by rewriting $\boldsymbol{R}^{(1)}$:

$$\boldsymbol{R}^{(1)} = \mathbf{M}\boldsymbol{R}^{(1)} \Leftrightarrow (\mathbf{I} - c\mathbf{A}^\top)\boldsymbol{R}^{(1)} = (c\boldsymbol{u}\boldsymbol{g}^\top + (1-c)\boldsymbol{u}\boldsymbol{e}^\top)\boldsymbol{R}^{(1)} \quad . \tag{7}$$

Looking at the right hand side we get:

$$(c\boldsymbol{u}\boldsymbol{g}^\top + (1-c)\boldsymbol{u}\boldsymbol{e}^\top)\boldsymbol{R}^{(1)} = \left(1 - c + c\sum \boldsymbol{R}_a^{(1)}\right)\boldsymbol{u} = \left(1 - c\sum \mathbf{A}^\top \boldsymbol{R}^{(1)}\right)\boldsymbol{u} \tag{8}$$

where $\boldsymbol{R}_a^{(1)}$ is the PageRank of all vertices with no outgoing edges. This gives:

$$\boldsymbol{R}^{(1)} = (\mathbf{I} - c\mathbf{A}^\top)^{-1} \left(1 - c \sum \mathbf{A}^\top \boldsymbol{R}^{(1)}\right) \boldsymbol{u} \ . \tag{9}$$

From Definition. 2 we have

$$\boldsymbol{R}^{(3)} = \frac{\boldsymbol{R}^{(1)}||\boldsymbol{v}||_1}{d} \Rightarrow \boldsymbol{R}^{(1)} = \frac{\boldsymbol{R}^{(3)} d}{||\boldsymbol{v}||_1} \ . \tag{10}$$

Substituting (10) into (9) gives:

$$\boldsymbol{R}^{(3)} = (\mathbf{I} - c\mathbf{A}^\top)^{-1} \frac{||\boldsymbol{v}||_1}{d} \left(1 - c \sum \mathbf{A}^\top \boldsymbol{R}^{(1)}\right) \boldsymbol{u} \ . \tag{11}$$

Since $d = 1 - \sum c\mathbf{A}^\top \boldsymbol{R}^{(1)}$ and $\boldsymbol{v} = ||\boldsymbol{v}||_1 \boldsymbol{u}$ we end up with

$$\boldsymbol{R}^{(3)} = (\mathbf{I} - c\mathbf{A}^\top)^{-1} \boldsymbol{v} \ . \tag{12}$$

Which is the same as what we got using Definition. 3. □

We will mainly use Definition. 3 while Definition. 2 is mainly there to show that it is very simple to go from $\boldsymbol{R}^{(1)}$ to $\boldsymbol{R}^{(3)}$ and to show that $\boldsymbol{R}^{(1)}$ and $\boldsymbol{R}^{(3)}$ are proportional, thus corresponds to the same ranking.

2 PageRank for Different Types of Vertices

Our aim is to use the topology of the graph in order to find an effective method to calculate PageRank by handling different types of vertices differently. We start by defining five different types of vertices.

Definition 4. For the vertices of a simple directed graph with no loops we define 5 distinct groups
$G_1, G_2, \ldots, G_5$

1. G_1: Vertices with no outgoing or incoming edges.
2. G_2: Vertices with no outgoing edges and at least one incoming edge (also called dangling nodes).
3. G_3: Vertices with at least one outgoing edge, but no incoming edges (also called root nodes).
4. G_4: Vertices with at least one outgoing and incoming edge, but which is not part of any directed cycle (no path from the vertex back to itself).
5. G_5: Vertices that is part of at least one cycle.

From the construction of the five groups it should be obvious that every vertex belong to a single group, with no vertices belonging to multiple or none at all. Note that Qing Yu et al [12] also divide the vertices into five groups however in a slightly different way. They group it by dangling and root nodes (G_2 and G_3), vertices that can be made into a dangling or root node by recursively removing dangling and root nodes (part of G_4) and remaining vertices (part of G_4 and G_5).

2.1 Vertices in G_1

Vertices in G_1 corresponds to the simplest possible vertices with no edges either to or from of the vertex. For these types of vertices we formulate the following theorem.

Theorem 2. *We let $e_g \in G_1$ and e_i be any other vertex, then*

$$\boldsymbol{R}^{(3)}_{e_g} = v_g, \text{ and } P(e_g \to e_i) = 0 \ . \tag{13}$$

Proof. We look at $\boldsymbol{R}^{(3)}_{e_g}$ as defined in Definition. 3. Since there obviously is no path from e_g back to itself we have

$$\sum_{k=0}^{\infty} (P(e_g \to e_g))^k = 1 \ . \tag{14}$$

Since no other node links to e_g, we have $P(e_i \to e_g) = 0$ for all e_i as well, and we end up with $\boldsymbol{R}^{(3)}_{e_g} = (0 + v_g)(1) = v_g$. Since e_g links to no other node it's obvious that $P(e_g \to e_i) = 0$ as well. □

The first part means that the PageRank $\boldsymbol{R}^{(3)}$ of vertices in G_1 is always equal to their weight in $\boldsymbol{v}$, while the second part means that the vertex have no influence on the PageRank of any other verticesx and can therefor simply be removed from the PageRank calculationof other vertices altogether. We note that we can effectively calculate the rank of vertices in G_1 whenever we want (before or after finding the rank of all other vertices).

2.2 Vertices in G_2

Vertices in G_2 are what is often called "dangling nodes" with the small addition that there is at least one other vertex linking to it (so as to not belong to G_1). It is already well known how to handle this type of vertices effectively as described in for example [1] where these vertices are first removed from the system, then PageRank is calculated for remaining vertices and last PageRank of these "dangling nodes" are found. We still take a short look at them here as well since it is very clear how to handle them using the definition of PageRank we use.

Theorem 3. *We let $e_g \in G_2$ and e_i be any other vertex, then*

$$\boldsymbol{R}^{(3)}_{e_g} = \sum_{e_i \in S,\ e_i \neq e_g} v_i P(e_i \to e_g) + v_g, \text{ and } P(e_g \to e_i) = 0 \ . \tag{15}$$

Proof. We look at $\boldsymbol{R}^{(3)}_{e_g}$ as defined in Definition. 3. Since there obviously is no path from e_g back to itself we have

$$\sum_{k=0}^{\infty} (P(e_g \to e_g))^k = 1 \ . \tag{16}$$

However since there is at least one incoming edge to e_g we have

$$\sum_{e_i \in S,\ e_i \neq e_g} v_i P(e_i \to e_g) > 0 \ . \tag{17}$$

Since e_g have no outgoing edges, it is obvious that $P(e_g \to e_i) = 0$ as well. □

From the second part it is obvious that the vertex $e_g \in G_2$ does not influence the PageRank of any other vertex and can therefor safely be removed from the graph for the PageRank calculation of all other vertices. We also see that since there is no path back to itself, if we know the PageRank of all other vertices we can calculate the PageRank of e_g. In fact since $\boldsymbol{R}^{(3)}_{e_j} = \sum_{i=1}^{n} c a_{ij} \boldsymbol{R}^{(3)}_{e_i} + v_j$ we can easily find the PageRank of vertices in G_2 as

$$\boldsymbol{R}^{(3)}_{e_g} = \sum_{i=1}^{n} c a_{ig} \boldsymbol{R}^{(3)}_{e_i} + v_g \tag{18}$$

which is a simple vector $\times$ vector multiplication. This means that as long as we have the PageRank of all vertices not in G_1 or G_2 we can then find the PageRank of those in G_2 very quickly as well. We note that in order to find the rank of the vertices in G_2 we first need to calculate the rank of all other vertices.

In practice this means that we start by dividing the whole system S in two parts $S = S_0 \cup S_d$ where S_d is the vertices in G_2 in S and S_0 contains all other vertices. This gives system matrix

$$c\mathbf{A} = \begin{bmatrix} c\mathbf{A}_0 & c\mathbf{A}_d \\ 0 & 0 \end{bmatrix} \tag{19}$$

where $\mathbf{A}_0$ is the part of $\mathbf{A}$ with edges to vertices not in G_2 and $\mathbf{A}_d$ contains the edges to vertices in G_2. In the same way we have the weight vector $\boldsymbol{V} = [\boldsymbol{v}_0, \boldsymbol{v}_d]$ where $\boldsymbol{v}_d$ is the part of the weight vector corresponding to vertices in G_2. Given the PageRank of the part without the vertices in G_2: $\boldsymbol{R}^{(3)}_{S_0}$ the PageRank of the vertices in G_2 is

$$\boldsymbol{R}^{(3)}_{S_d} = c\mathbf{A}_d^{\top} \cdot \boldsymbol{R}^{(3)}_{S_0} + \boldsymbol{v}_d \ . \tag{20}$$

It is also possible to repeatedly remove vertices in G_2 until no longer possible and then apply the above procedure multiple times, once for every set of vertices removed.

2.3 Vertices in G_3

For vertices in G_3 we get something similar as for those in G_2.

Theorem 4. *We let $e_g \in G_3$ and e_i be any other vertex, then*

$$\boldsymbol{R}^{(3)}_{e_g} = v_g, \ \textit{and} \sum_{e_i \in S,\ e_i \neq e_g} P(e_g \to e_i) > 0 \ . \tag{21}$$

Proof. Again we look at $\boldsymbol{R}_{e_g}^{(3)}$ as defined in Definition. 3. Since there obviously is no path from e_g back to itself we have

$$\sum_{k=0}^{\infty} (P(e_g \to e_g))^k = 1 \ . \tag{22}$$

Since there is no incoming edges to e_g, we have $P(e_i \to e_g) = 0$ for all e_i as well, and we end up with $\boldsymbol{R}_{e_g}^{(3)} = (0 + v_g)(1) = v_g$. Since e_g has outgoing edges there is at least a path from e_g to the vertices e_g links to

$$\sum_{e_i \in S,\ e_i \neq e_g} P(e_g \to e_i) > 0 \ . \tag{23}$$

□

While vertices in G_2 did not change the PageRank of other vertices and could be calculated after other vertices effectively, vertices in G_3 like those in G_1 can easily be calculated before all other vertices. However they cannot be ignored when calculating the PageRank of remaining vertices. We take a closer look at calculating PageRank of other vertices given the Pagerank of a vertex $e_g \in G_3$.

Theorem 5. *Given $\boldsymbol{R}_{e_g}^{(3)} = v_g$, $e_g \in G_3$. We can write the PageRank of another general vertex e_i as*

$$\boldsymbol{R}_{e_i}^{(3)} = \left(v_i + v_g ca_{gi} + \sum_{\substack{e_j \in S \\ e_j \neq e_i, e_g}} (v_j + v_g ca_{gj}) P(e_j \to e_i) \right) \left(\sum_{k=0}^{\infty} (P(e_i \to e_i))^k \right) \tag{24}$$

where ca_{gi} is the one-step probability to go from e_g to e_i.

Proof. Again we look at $\boldsymbol{R}_{e_g}^{(3)}$ as defined in Definition. 3. Since we know that there is no path from e_i back to e_g we know that the right hand side will be identical for all other vertices. We rewrite the influence of e_g using

$$v_g P(e_g \to e_i) = v_g ca_{gi} + \sum_{\substack{e_j \in S \\ e_j \neq e_i, e_g}} v_g ca_{gj} P(e_j \to e_i) \ . \tag{25}$$

We can now rewrite the left sum in Definition. 3:

$$\sum_{e_i \in S, e_i \neq e_j} v_i P(e_i \to e_j) = v_g ca_{gi} + \sum_{\substack{e_j \in S \\ e_j \neq e_i, e_g}} (v_j + v_g ca_{gj}) P(e_j \to e_i) \tag{26}$$

which when substituted into (3) proves the theorem. □

It is clear that while e_g influences the PageRank of other vertices, if we change the weights of the vertices linked to by $e_g \in G_3$ we can remove e_g when calculating the PageRank of other vertices. In a way it is similar to the vertices

in G_2 but rather instead of first computing the main part and then compute PageRank of the vertices in G_2 we first compute the PageRank of vertices in G_3 and then the main part of the vertices.

Using Theorem 4 we can find the PageRank of any vertex in G_3 as the weight $\boldsymbol{v}_i$ for that vertex. Likewise it is also simple to find the rank of any vertices whose incoming edges are only from vertices in G_3.

To find the PageRank of the remaining vertices we first need to modify their initial weights to accommodate for the PageRank of the vertices in G_3 and then calculate PageRank with the new weight vector. We recall that PageRank can be seen as a sum of all probabilities of ending up in a vertex when starting in every vertex once with an initial "probability" equal to its weight.

This means that we only need to change the weights for the vertices that are linked to by the vertices in G_3 and not any other. We get the new weights in the same way as when we calculated the PageRank of the vertices with incoming edges from only vertices in G_3:

$$v_{i,new} = \sum_{e_j \in G_3} ca_{ji} \boldsymbol{R}_j^{(3)} + v_i \ . \tag{27}$$

Now that we have the new weights it's a simple matter of applying the PageRank algorithm of choice on the remaining system and get their PageRank.

2.4 Vertices in G_4

Vertices in G_4 can be seen as transient states, once you leave them you can never get back. But compared to G_2 and G_3 there are edges both to and from of the vertex. While not as simple as for the previous groups we will see that vertices in G_4 can sometimes change group to one of G_2 or G_3 when a vertex in G_2 or G_3 is removed from the graph.

Theorem 6. *We let $e_g \in G_4$ and e_i be any other vertex, then*

$$\boldsymbol{R}_{e_g}^{(3)} = \sum_{e_i \in S,\ e_i \neq e_g} v_i P(e_i \to e_g) + v_g, \ \textit{and} \sum_{e_i \in S,\ e_i \neq e_g} P(e_g \to e_i) > 0 \ . \tag{28}$$

Proof. Again we look at $\boldsymbol{R}_{e_g}^{(3)}$ as defined in Definition. 3. Since there is no path from e_g back to itself (e_g is not part of any cycle) we have $\sum_{k=0}^{\infty} \left(P(e_g \to e_g)\right)^k = 1$. However since there is at least one incoming edge to e_g and e_g links to at least one other vertex we have

$$\sum_{e_i \in S,\ e_i \neq e_g} v_i P(e_i \to e_g) > 0, \text{ and } \sum_{e_i \in S,\ e_i \neq e_g} P(e_g \to e_i) > 0 \ . \tag{29}$$

□

Similar to G_2 we see that if we know the PageRank of all other vertices in G_3, G_4, G_5 we can easily find the PageRank of one vertex in G_4 as well. Similarly if we know the PageRank of e_g we can remove e_g from the graph by changing weights as we did for vertices in G_3. This might look like a problem

in that we would need the PageRank of the other vertices to increase the speed of calculating PageRank of these same vertices. However since $e_g \in G_4$ is not part of any cycle, the PageRank of any vertex for which can be reached from e_g does not influence the PageRank of any vertex which can reach e_g. Vertices for which there is no path in either direction with e_g obviously doesn't effect nor are affected by the PageRank of e_g. The obvious practical use of these vertices is the fact that they can change from G_4 to either of G_2 or G_3 when removing vertices in G_2 or G_3

Not all vertices in G_4 can change in this way though as in the case when there are vertices in G_5 with paths both from and to e_g. In this case more information is needed.

2.5 Vertices in G_5

The existence of vertices in G_5 is the reason we need methods such as the Power method. We will not go deeper into if it is possible to divide G_5 itself into more groups where some might have more effective calculation methods. For vertices in G_5 we can make no obvious simplification as for the other groups. We see that the biggest obstacle in how fast we can find the PageRank (theoretically) depends on the existence and size of cycles in the graph rather than tree structures.

2.6 A Simple Scheme to Calculate PageRank

We are now ready to give a simple scheme for calculating PageRank in which we combine what we have found this far to calculate PageRank. This method is very similar to the one proposed by Qing et al [12] apart from the first step made possible because of how we define PageRank.

1. If possible split up the system into multiple disjoint parts $S_0, S_1, S_2, \ldots$ and calculate PageRank for every subsystem individually.
2. Repeatedly remove all vertices in G_2 and any new vertices in G_2 created until there are no longer any vertices in G_2 left.
3. Remove vertices in G_3 in the same way, what is left we call the skeleton.
4. Calculate PageRank of all vertices in G_3 as discussed in Sect. 2.3
5. Change the weight vector for the skeleton to accommodate the influence of the vertices in G_3 as discussed in Sect. 2.3.
6. Calculate PageRank of the skeleton using any method of choice.
7. Calculate the PageRank of the vertices in G_2 using the PageRank of the skeleton and vertices in G_3 as discussed in Sect. 2.2.

This could potentially be made faster by improving the main PageRank calculation in (6) by for example an adaptive PageRank algorithm [10] or by systematically aggregating the vertices as in [7].

3 PageRank for Different Strongly Connected Components

While removing vertices as discussed earlier can give a significant improvement when calculating PageRank, there is one problem in that a large amount of the vertices in G_4 could lie between different strongly connected components (henceforth only written as "component"), and thus will not be removed according to the earlier scheme. To guarantee that we can remove all of these as well as give a natural way to divide the problem in smaller problems (for parallel computing for example) we will take a look at the topology of the graph.

By first finding all components of the graph and their topological ordering for example by using Tarjan's algorithm [11]. By regarding every component as a single vertex we can partition them in the previously found groups $G_1, \dots G_5$. The rank of these components can be handles in a way similar to that of the single vertices in the previous section.

Component in G_1. As with a vertex in G_1, PageRank of a connected component with no edges to or from any other connected component can be calculated by itself at any time during the calculations according to the scheme in Sect. 2.6.

Component in G_2. As with a vertex in G_2, PageRank of a component with no edges to any other component does not influence the PageRank of any other component. Similarly we can find the PageRank of the component itself by first calculating for all other components leading to it and then change initial weight vector as we did when considering a vertex in G_3 in Sect. 2.3.

Component in G_3. Once again we can compare it with a single vertex in the same group, since the component have no incoming edges from any other components its PageRank we can calculate its PageRank before any other. In the same way as for single vertices we need to adjust the weights for vertices outside the component linked to by any edge in the component.

Component in G_4 and G_5. Since G_5 will be empty (otherwise they would compromise a single larger component), all components in G_4 will eventually be reduced to one in G_3 if we repeatedly calculate the PageRank of all components in G_3, adjust the weight for all other vertices and remove them from the graph.

3.1 A New Scheme to Calculate PageRank

This gives us a new scheme with which to calculate PageRank:

1. Find components of the graph and their topological order. Optionally merge 1-vertex components.
2. Calculate PageRank of all components in G_3 using the scheme in Sect. 2.6.
3. Adjust the weight vector according to the PageRank of the components we just calculated and remove them from the graph.

4. Repeat step 2-3 until there is no components in G_3 left.
5. Calculate PageRank of remaining components (in G_1).

Note that PageRank for components in G_1 can be calculated at any time. This second method depend on a fast way to find components in the graph which may or may not be worth it depending on the structure and size of the graph as well as the error tolerance used. There is also the possibility that the extra overhead needed compared to the previous algorithm, both in finding the components as well as the handling of them during the PageRank steps themselves could influence the performance of the algorithm negatively.

One advantage when partitioning the graph into multiple components is that it makes it possible to choose the most suitable algorithm for every component rather than one method for the whole graph.

3.2 Small components

While the numerical methods used to calculate PageRank for one component is done in linear time in the number of edges, the number of iterations needed is generally not that much smaller than for a very large component. Because of this it is often more suitable to solve the equation system analytically if the graph is small, say 10 vertices or less.

Using equation 6 we get the equation system $(\mathbf{I} - c\mathbf{A}^\top)\boldsymbol{R}^{(3)} = \boldsymbol{v}$ which we already know to have a unique solution found using any standard method.

Calculating PageRank for these components analytically have the added benefit that no extra error is introduced from the method itself, while the numerical methods will always have a small error. Regardless it should be noted that even if a component is calculated analytically, there could still be other components with edges targeting vertices in this components, hence the initial weight v could have some small errors (decided by the error tolerance for the numerical methods).

Iin many real networks such as a graph over the Web there is often one very large components but also a large number of very small components. Even if these components are very small components, because of the great number of them they signify a significant amount of the total number of vertices and time needed.

3.3 Acyclic subgraphs

If 1-vertex components are merged in step 1, then this corresponds to a subgraph with no cycles. The algorithm for acyclic subgraphs exploits the fact that there are no cycles to calculate the PageRank of the graph using a variation of Breadth first search. Since An acyclic subgraph have no vertices in G_5 the value in the right hand paranthesis in 3 is equal to one for all vertices, thus the rank of a vertex v can be decided from the initial weight and the rank of all vertices with incomming edges to v in the same way as we did for those in G_3.

The algorithm start by calculating the in-degree of every vertex and store it in `v.degree` for each vertex v, this can be done by looping through all edges

in the subgraph once. We also keep a value `v.rank` initialized to corresponding value in the weight vector for each vertex. The breadth first search itself can be described by the following psuedocode.

```
for each vertex v
   if v.degree == 0
    Queue.enqueue(v)
    while Queue.size > 0
      w = Queue.dequeue()
      for each edge (w,u)
        u.rank = u.rank+w.rank*W(w,u)
        u.degree--
        if (w.degree == 0) Queue.enqueue(w)
      end
      w.degree--
    end
  end
end
```

Here `W(w,u)` is the weight on the (w,u)-edge. The difference between this and ordinary BFS is that we only add a vertex v to the queue when we have visited all incomming edges to v rather than when we have visited the first incomming edge as for ordinary breadth first search. We also loop through all edges to ensure that we wisit each vertex once since there could be multiple initial vertices which have a zero in-degree.

Looking at the time complexity it is easy to see that we visit every vertex and every edge once doing constant time work, hence we have the same time complexity as for ordinary breadth first search $O(|V|+|E|) \approx O(|E|)$, if $|E| \gg |V|$. While it has the same time complexity as most numerical methods to calculate PageRank, in practice it will often be much faster in that the coefficient in front will be smaller, especially as the error tolerance of the numerical method decreases.

4 Verifying the linear time complexity of the algorithm

Consider a graph G with E edges and V vertices, it is easy to show that the ordinary PageRank algorithm can be done in $O(E)$. For example using the power method (by iterating $\boldsymbol{R^{(1)}}_{n+1} = c\mathbf{A}^\top \boldsymbol{R^{(1)}}_n + (1 - \sum c\mathbf{A}^\top \boldsymbol{R}^{(1)}_n)\boldsymbol{u}$) we need to do one matrix$\times$vector multiplication with a sparse matrix needing $O(E)$ operations and one vector addition needing $O(V)$ operations. In addition some kind of convergence criterion need to be used (typically $O(V)$, such as the max or mean difference between iterations). The number of iterations does not depend on the number of vertices (at least not directly), but primary on the convergence criterion used.

Finding strongly connected components and their topological order in step 1 can be done in $O(E)$ time using a depth first search such as Tarjan's algorithm [11]. While both this and the PageRank algorithm itself is linear in the number

of edges, the depth first search is generally much faster since it only needs to access every edge once, while the Power method needs to access every edge once in every iteration.

The work done in step 2 and step 5 is the same as for the original Power method (or whichever method is used) except that edges between components are not used $O(E_{scc})$, where E_{scc} is the number of edges within endpoints within the same strongly connectd component.

In step 3 consists of a single matrix×vector multiplication similar to one step in the power method for each set of G_3 components. In total over all components each edge between two strongly connected components is used exactly once, this gives $O(E_b)$ where E_b denotes the number of edges between components. This gives a total time complexity of $O(E_{scc} + E_b) = O(E)$ for step 2 and 3, although it should be noted that the work needed for edges between components is much less since they are only needed once.

Thus we can see that we have the same time complexity overall for the the method $O(E)$ however some edges are only needed once in the PageRank step compared to those that might be needed tens or hundreds of times depending on graph structure and error tolerance. Unfortunately there is also an extra cost in step one needed in order to find the strongly connected components and order them appropriately.

5 Conclusions and future work

We have seen how to theoretically improve the computation of PageRank using the structure of the graph, both for different types of vertices as well as for entire strongly connected components in the graph. We could also see that we could use the same structures to potentially divide the problem into many smaller problems potentially making it easier to implement in parallel on multiple computers as well as allowing for the use of sepererate methods for different components. We have not made any assumptions about the structure of the graph, however it is obvious that calculating for certain types of graphs we would gain a lot more from what we have looked at here than others. Especially we see that the presence of long cycles limits the possibility for us to divide the graph into smaller parts, both on the vertex level as well as on the component level.

The next step which is underway is implementing the method (along with a way to combine certain 1-vertex components) and comparison with previous methods. Other remaining things to look at is if certain strongly connected components could be handles more effectively. A start may be to look at vertices with only one incoming or outgoing edge.

References

1. F. Andersson and S. Silvestrov. The mathematics of internet search engines. *Acta Appl. Math.*, 104:211–242, 2008.
2. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.

3. K. Bryan and T. Leise. The $25,000,000,000 eigenvector: The linear algebra behind google. *SIAM Review*, 48(3):569–581, 2006.
4. C. Engström. Pagerank as a solution to a linear system, pagerank in changing systems and non-normalized versions of pagerank. Master's thesis, Mathematics, Centre for Mathematical sciences, Lund Institute of Technology, Lund University, May 2011:E31. LUTFMA-3220-2011.
5. C. Engström, S. Silvestrov, and T. Hamon. Variations of pagerank with application to linguistic data. In *Applied Stochastic Models and Data Analysis (ASMDA 2013). The 15th Conference of the ASMDA International Society*, pages E–I, 9–18, 2013.
6. T. Haveliwala and S. Kamvar. The second eigenvalue of the google matrix. Technical Report 2003-20, Stanford InfoLab, 2003.
7. H. Ishii, R. Tempo, E.-W. Bai, and F. Dabbene. Distributed randomized pagerank computation based on web aggregation. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 3026–3031, 2009.
8. S. Kamvar and T. Haveliwala. The condition number of the pagerank problem. Technical Report 2003-36, Stanford InfoLab, June 2003.
9. C. P. Lee, G. H. Golub, and S. A. Zenios. A two-stage algorithm for computing pagerank and multistage generalizations. *Internet Mathematics*, 4(4):299–327, 2007.
10. K. Sepandar, H. Taher, and G. Gene. Adaptive methods for the computation of pagerank. *Linear Algebra and its Applications*, 386(0):51 – 65, 2004.
11. R. E. Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.
12. Q. Yu, Z. Miao, G. Wu, and Y. Wei. Lumping algorithms for computing google's pagerank and its derivative, with attention to unreferenced nodes. *Information Retrieval*, 15(6):503–526, 2012.

A vine and gluing copula model for permeability stochastic simulation

Arturo Erdely[1] and Martin Diaz-Viera[2]

[1] Actuarial Science Program, Facultad de Estudios Superiores Acatlan, Universidad Nacional Autonoma de Mexico, Mexico
(E-mail: arturo.erdely@comunidad.unam.mx)
[2] Gerencia de Ingenieria de Recuperacion Adicional, Instituto Mexicano del Petroleo, Mexico
(E-mail: mdiazv@imp.mx)

Abstract. Statistical dependence between petrophysical properties in heterogeneous formations is usually nonlinear and complex; therefore, traditional statistical methods based on assumptions of linearity and normality are usually not appropriate. Copula based models have been previously applied to this kind of variables but it seems to be very restrictive to find a single copula family to be flexible enough to model complex dependencies in highly heterogeneous porous media. The present work combines vine copula modeling with a bivariate gluing copula approach to model rock permeability using vugular porosity and measured P-wave velocity as covariates in a carbonate double-porosity formation at well log scale.
Keywords: Vine and gluing copulas, nonlinear dependence, petrophysical modeling.

1 Copula basics

A *copula function* is the functional link between the joint probability distribution function of a random vector and the marginal distribution functions of the random variables involved. For example, in a bivariate case, if (X, Y) is a random vector with joint probability distribution $F_{XY}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ with continuous marginal distribution functions F_X and F_Y then by *Sklar's Theorem*[19] there exists a unique bivariate copula function $C_{XY} : [0, 1]^2 \to [0, 1]$ such that $F_{XY}(x, y) = C_{XY}(F_X(x), F_Y(y))$. Therefore, all the information about the dependence between X and Y is contained in the underlying copula C_{XY}, since F_X and F_Y only explain the individual (marginal) behavior of such random variables. As an example, for continuous random variables, X and Y are independent if and only if $C_{XY}(u, v) = \Pi(u, v) := uv$.

As a consequence of results by Hoeffding[9] and Fréchet[6], particularly what is known as the *Fréchet-Hoeffding bounds* for joint probability distribution functions, Sklar's Theorem leads to the following sharp bounds for any bivariate copula: $W(u, v) \leq C_{XY}(u, v) \leq M(u, v)$ for all u, v in $[0, 1]$, where $W(u, v) := \max\{u+v-1, 0\}$ and $M(u, v) := \min\{u, v\}$ are themselves copulas. W (respectively M) is the underlying copula of a bivariate random vector of

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

continuous random variables (X, Y) if, say, Y is an almost surely decreasing (respectively increasing) function of X.

Formal definitions and main properties of copula functions are covered in detail in Nelsen[14] and Durante and Sempi[3]. Among many other properties, any copula C is a uniformly continuous function, and in particular its *diagonal section* $\delta_C(t) := C(t,t)$ is uniformly continuous and nondecreasing on $[0,1]$. In terms of the Fréchet-Hoeffding bounds, we get $\max\{2t-1, 0\} \leq \delta_C(t) \leq t$ for all t in $[0,1]$.

Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ denote an observed sample of size n from a bivariate random vector (X, Y) of continuous random variables. We may estimate the underlying copula C_{XY} by the *empirical copula* C_n, see Deheuvels[2], which is a function with domain $\{\frac{i}{n} : i = 0, 1, \ldots, n\}^2$ defined as:

$$C_n\left(\frac{i}{n}\,,\,\frac{j}{n}\right) := \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{rank(x_k) \leq i\,,\, rank(y_k) \leq j\} \tag{1}$$

and its convergence to the true copula C_{XY} has also been proved, see Rüschendorf[17] and Fermanian *et al.*[5]. Strictly speaking, the empirical copula is not a copula since it is only defined on a finite grid, but by Sklar's Theorem C_n may be extended to a copula. Based on the empirical copula several goodness-of-fit tests have been developed, see for example Genest *et al.*[7], to choose the best parametric family of copulas from an already existing long catalog, see for example chapter 4 in Joe[11].

The underlying copula C_{XY} is invariant under strictly increasing transformations of X and Y, that is $C_{XY} = C_{\alpha(X),\beta(Y)}$ for any strictly increasing functions α and β. Recall that for any continuous random variable X we have that the random variable $F_X(X)$ is uniformly distributed on the open interval $]0,1[$. Let $U := F_X(X)$ and $V := F_Y(Y)$, then (X, Y) has the same underlying copula as (U, V) and by Sklar's Theorem $F_{UV}(u, v) = C_{UV}(F_U(u), F_V(v)) = C_{UV}(u, v)$. So the transformed sample $\{(u_1, v_1), \ldots, (u_n, v_n)\}$ where $(u_k, v_k) = (F_X(x_k), F_Y(y_k))$ may be considered as *observations* from the underlying copula C_{XY}. If F_X and F_Y are unknown (which is usually the case) they can be replaced by the empirical aproximation $F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{x_k \leq x\}$ and in such case we obtain what is known as *pseudo-observations* of the underlying copula C_{XY}, which are used for copula estimation purposes, since they are equivalent to the ranks in (1).

2 Gluing copulas

Sklar's Theorem is also useful for building new multivariate probability models. For example, if F and G are univariate probability distribution functions, and C is any bivariate copula, then $H(x, y) := C(F(x), G(y))$ defines a joint probability distribution function with univariate marginal distributions F and G. Several methods for constructing families of copulas have been developed (geometric methods, archimedean generators, ordinal sums, convex sums, shuffles) and among them we may include *gluing copulas* by Siburg and Stoimenov[18], which we will illustrate in a very particular case: let C_1 and C_2 be two given

bivariate copulas, and $0 < \theta < 1$ a fixed value, we may scale C_1 to $[0, \theta] \times [0, 1]$ and C_2 to $[\theta, 1] \times [0, 1]$ and *glue* them into a single copula:

$$C_{1,2,\theta}(u, v) := \begin{cases} \theta C_1(\frac{u}{\theta}, v), & 0 \leq u \leq \theta, \\ (1-\theta)C_2(\frac{u-\theta}{1-\theta}, v) + \theta v, & \theta \leq u \leq 1. \end{cases} \tag{2}$$

A gluing copula construction may easily lead to a copula with a diagonal section $\delta_{1,2,\theta}(t) = C_{1,2,\theta}(t, t)$ that has a discontinuiy in its derivative at the *gluing point* $t = \theta$. This fact may be taken into consideration when trying to fit a parametric copula to observed data, since common families of copulas have diagonal sections without discontinuities in their derivatives, and if the *empirical diagonal* $\delta_n(\frac{i}{n}) := C_n(\frac{i}{n}, \frac{i}{n})$ strongly suggests there is one or more points at which a discontinuity of the derivative occurs, an appropriate data partition by means of finding some gluing points could be helpful to model the underlying copula by the gluing copula technique.

For a more specific example, in the particular case $C_1 = M$ and $C_2 = \Pi$ it is straightforward to verify that for $0 \leq t \leq \theta$ we get a diagonal section $\delta_{1,2,\theta}(t) = \theta t$, while for $\theta \leq t \leq 1$ we get $\delta_{1,2,\theta}(t) = t^2$ and clearly the left and right derivatives at the gluing point $t = \theta$ are not the same, see Figure 1.

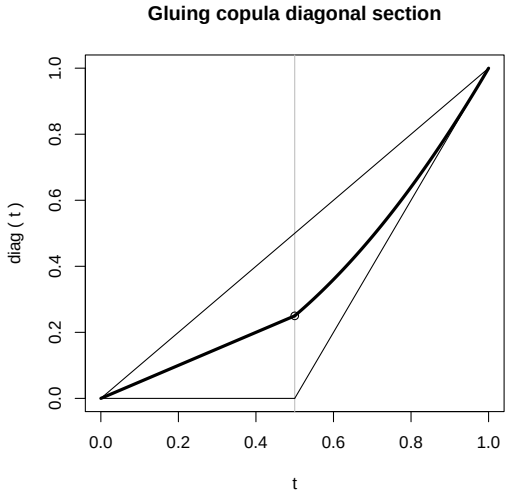

Fig. 1. Diagonal section of the resulting gluing copula with $C_1 = M$, $C_2 = \Pi$ and gluing point $\theta = \frac{1}{2}$

3 Trivariate vine copulas

In the previous sections we summarized some main facts about bivariate copulas, but Sklar's Theorem is valid for any $d \geq 2$ random variables. For example,

in the case of a trivariate random vector (X_1, X_2, X_3) of continuous random variables with joint probability distribution $F_{123}(x_1, x_2, x_3) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, X_3 \leq x_3)$ and marginal univariate distributions F_1, F_2, and F_3, by Sklar's Therem there exists a unique underlying copula $C_{123} : [0,1]^3 \to [0,1]$ such that $F_{123}(x_1, x_2, x_3) = C_{123}(F_1(x_1), F_2(x_2), F_3(x_3))$. In case F_{123} is *absolutely continuous* we may obtain the following expression for the trivariate joint density:

$$f_{123}(x_1, x_2, x_3) = c_{123}(F_1(x_1), F_2(x_2), F_3(x_3)) f_1(x_1) f_2(x_2) f_3(x_3) \tag{3}$$

where the copula density $c_{123}(u, v, w) = \frac{\partial^3}{\partial u \partial v \partial w} C_{123}(u, v, w)$ and the marginal densities $f_k(x) = \frac{d}{dx} F_k(x)$, $k \in \{1, 2, 3\}$. According to Kurowicka[13]:

> The choice of copula is an important question as this can affect the results significantly. In the bivariate case $[d = 2]$, this choice is based on statistical tests when joint data are available [...] Bivariate copulae are well studied, understood and applied [...] Multivariate copulae $[d \geq 3]$ are often limited in the range of dependence structures that they can handle [...] Graphical models with bivariate copulae as building blocks have recently become the tool of choice in dependence modeling.

The main idea behind *vine copulas* (or pair-copula constructions) is to express aribitrary dimensional dependence structures in terms of bivariate copulas and univariate marginals. For example, we may rewrite the trivariate joint density (3) in the following manner by conditioning in one of the random variables, say X_1 :

$$\begin{aligned} f_{123} &= f_{23|1} \cdot f_1 \\ &= c_{23|1}(F_{2|1}, F_{3|1}) \cdot f_{2|1} \cdot f_{3|1} \cdot f_1 \\ &= c_{23|1}(F_{2|1}, F_{3|1}) \cdot \frac{f_{12}}{f_1} \cdot \frac{f_{13}}{f_1} \cdot f_1 \\ &= c_{23|1}(F_{2|1}, F_{3|1}) \cdot c_{12}(F_1, F_2) \cdot c_{13}(F_1, F_3) \cdot f_1 \cdot f_2 \cdot f_3 \end{aligned} \tag{4}$$

with other two similar possibilities by conditioning on random variables X_2 or X_3. If $\{(x_{1k}, x_{2k}, x_{3k})\}_{k=1}^n$ is a an observed sample size n from an absolutely continuous random vector (X_1, X_2, X_3) we may use the bivariate observations $\{(x_{1k}, x_{2k}\}_{k=1}^n$ to estimate c_{12} and $F_{2|1}$, and we use $\{(x_{1k}, x_{3k}\}_{k=1}^n$ to estimate c_{13} and $F_{3|1}$. Following the ideas in Gijbels *et al.*[8] we obtain the following expression for the conditional bivariate joint distribution of (X_2, X_3) given $X_1 = x_1$:

$$\begin{aligned} F_{23|1}(x_2, x_3 \,|\, x_1) &= \mathbb{P}(X_2 \leq x_2, X_3 \leq x_3 \,|\, X_1 = x_1) \\ &= C_{23|1}\big(F_{2|1}(x_2 \,|\, x_1), F_{3|1}(x_3 \,|\, x_1) \,\big|\, x_1\big) \end{aligned} \tag{5}$$

Here the value x_1 becomes a parameter for the conditional bivariate copula $C_{23|1}$ and for the conditional univariate marginals $F_{2|1}$ and $F_{3|1}$. In case there is some kind of evidence (empirical or expert-based) to assume that the underlying bivariate copula for $F_{23|1}$ does not depend on the value of the conditioning variable, we have what is known as a *simplifying assumption*, see for

example Joe[11], and so to estimate such bivariate copula $C^*_{23} \equiv C_{23|1}$ again we may follow the ideas in Gijbels *et al.*[8] and use the pseudo-observations $\{(u_{2k}, u_{3k}) = \big(F_{2|1}(x_{2k} \,|\, x_{1k}), F_{3|1}(x_{3k} \,|\, x_{1k})\big)\}_{k=1}^{n}$.

4 Application to petrophysical data

As mentioned in Erdely and Diaz-Viera[4]:

> Assessment of rock formation permeability is a complex and challenging problem that plays a key role in oil reservoir modeling, production forecast, and the optimal exploitation management [...] Dependence relationships [among] petrophysical random variables [...] are usually nonlinear and complex, and therefore those statistical tools that rely on assumptions of linearity and/or normality and/or existence of moments are commonly not suitable in this case.

In the present work we apply a trivariate vine copula model to petrophysical data from Kazatchenko *et al.*[12] for variables X_1 = vugular porosity (PHIV), X_2 = measured P-wave velocity (VP), and X_3 = permeability(K), see Figure 2 for bivariate scatterplots and bivariate copula pseudo-observations.

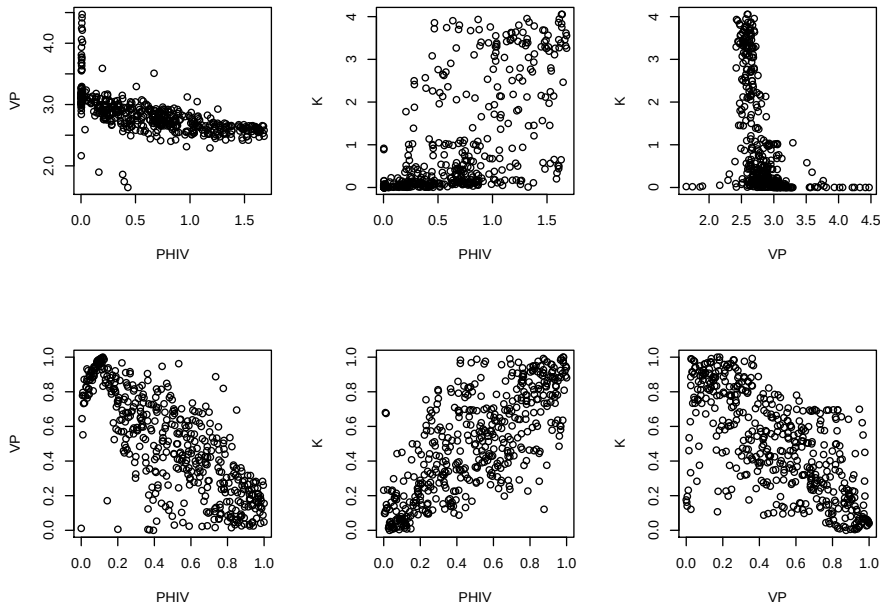

Fig. 2. *First row:* bivariate scatterplots. *Second row:* bivariate copula pseudo-observations.

First we searched for empirical evidence to check if a simplifying assumption is reasonable by splitting the pseudo-observations

$\{(u_{2k}, u_{3k}) = \big(F_{2|1}(x_{2k} \,|\, x_{1k}), F_{3|1}(x_{3k} \,|\, x_{1k})\big)\}_{k=1}^{n}$ in two sets A and B depending on whether the conditioning variable was less or greater than its median, and use them for an equality of copulas hypothesis test $\mathcal{H}_0 : C_A = C_B$ by Rémillard and Scaillet[16] implemented in the `TwoCop` R-package[15], see Table 1 for a summary of the results obtained. An extremely low p-value leads to the conclusion of rejecting a simplifying assumption, since lower values of the conditioning variable suggest a different dependence structure that the one corresponding to higher values. From Table 1 we conclude that a simplfying assumption condtioning on variable X_3 is definitely rejected, and conditioning on X_1 would be the best option in this case.

Conditioning variable	Simplifying assumption p-value
X_1	0.34
X_2	0.13
X_3	0.00

Table 1. p-values from Rémillard-Scaillet test adapted to test for simplifying assumption.

For the three bivariate copulas needed in the trivariate vine copula model (4) no single family of parametric bivariate copulas was able to achieve an acceptable goodness-of-fit, according to results obtained with the `copula` R-package[10]. Therefore a gluing approach has been applied, using a heuristic procedure to find gluing point candidates, called also *knots*, for a piecewise cubic polynomial fit (a particular case of *splines*) to the empirical diagonal δ_n but without the usual assumption of having continuous first or second derivative at the knots, since for gluing copula purposes that is exactly what we are looking for: points of discontinuity in the derivative of the diagonal section of the underlying copula.

Let $\mathcal{K} := \{t_0, \ldots, t_m\}$ be a set of $m+1$ knots in the interval $[0,1]$ such that $0 = t_0 < t_1 < \cdots < t_m = 1$. Consider the set $\mathcal{P}$ of all continuous functions p on $[0,1]$ such that:

- $p(t_i) = \delta_n(t_i)$, $i \in \{0, 1, \ldots, m\}$
- p is a cubic polynomial on $[t_{i-1}, t_i]$, $i \in \{1, \ldots, m\}$

The goal is to find the smallest sets of knots $\mathcal{K}$ such that the *mean squared error (MSE)* of piecewise polynomial approximations to each empirical diagonal δ_n is minimal and such that it is possible to reach an acceptable goodness-of-fit of bivariate copulas for the data partitions induced by each $\mathcal{K}$:

Step 1 Calculate pseudo-observations $\mathcal{S} := \{(u_k, v_k) : k = 1, \ldots, n\}$ and rearrange pairs such that $u_1 < \cdots < u_n$.

Step 2 Calculate empirical diagonal $\mathcal{D}_n := \{(\frac{i}{n}, \delta_n(\frac{i}{n})) : i = 0, 1, \ldots, n\}$.

Step 3 Find optimal knot (or gluing point) $t^* = \frac{i^*}{n}$ such that $\mathcal{K} = \{0, t^*, 1\}$ leads to minimal MSE on $\mathcal{D}_n$.

Step 4 Define subsets $\mathcal{G}_1$ and $\mathcal{G}_2$ from $\mathcal{S}$ such that $\mathcal{G}_1 := \{(u_k, v_k) \in \mathcal{S} : u_k \leq t^*\}$ and $\mathcal{G}_2 := \{(u_k, v_k) \in \mathcal{S} : u_k \geq t^*\}$.

Step 5 Apply goodnes-of-fit tests for parametric copulas in each subset $\mathcal{G}_1$ and $\mathcal{G}_2$.

Step 6 If an acceptable fit is reached in both subsets, we are done. Otherwise, apply steps 1 to 5 to the subset(s) which could not fit.

In Table 2 we present a summary of results, specifying how many partitions were needed and the best copula goodness-of-fit achieved on each one, for each bivariate relationship required by (4), making use of the `copula` R-package[10].

Bivariate dependence	Best parametric copula fit	p-value
X_1 , $-X_2$	Plackett*	0.6079
	Galambos*	0.1384
	Plackett	0.3941
	independence	0.5200
X_1 , X_3	Plackett*	0.6539
	Clayton	0.1494
	Husler-Reiss	0.8586
$-X_2$, $X_3 \mid X_1$	Plackett*	0.3541
	Clayton*	0.4800

Table 2. Families of copulas indicated with * means that the transformed copula $C^*(u, v) = u + v - 1 + C(1 - u, 1 - v)$ was used, where C is the original copula family.

5 Final remark

According to Czado and Stöber[1]:

> [...] compared to to the scarceness of work on multivariate copulas, there is an extensive literature on bivariate copulas and their properties. Pair copula constructions (PCCs) build high-dimensional copulas out of bivariate ones, thus exploiting the richness of the class of bivariate copulas and providing a flexible and convenient way to extend the bivariate theory to arbitrary dimensions.

But even expecting a single copula family to be able to model a complex bivariate dependency seems to be still too restrictive, at least for the petrophysical variables under consideration in this work. In such case, an alternative found was to apply a gluing copula approach[18]: decomposing bivariate samples into subsamples whose dependence structures were simpler to model by known parametric families of copulas, taking advantage of already existing tools (and their computational implementations) for bivariate copula estimation.

Acknowledgement

The present work was supported by project IN115914 from Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) at Universidad Nacional Autónoma de México.

References

1. C. Czado and J. Stöber. Pair Copula Constructions. In J.-F. Mai and M. Scherer, editors, em Simulating Copulas, pp. 185–230, Imperial College Press, London, 2012.
2. P. Deheuvels. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci.*, 65, 5, 274–292, 1979.
3. F. Durante and C. Sempi. *Principles of Copula Theory*, CRC Press, Boca Raton, 2016.
4. A. Erdely and M. Diaz-Viera. Nonparametric and Semiparamegric Bivariate Modeling of Petrophysical Porosity-Permeability Dependence from Well Log Data. In P. Jaworski, F. Durante, W. Härdle, T. Rychlik, editors, *Copula Theory and Its Applications*, pp. 267–278, Springer-Verlag, Berlin Heidelberg, 2010.
5. J-D. Fermanian, D. Radulovíc, M. Wegcamp. Weak convergence of empirical copula processes. *Bernoulli*, 10, 847–860, 2004.
6. M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon*, 14, (Sect. A Ser.3), 53–77, 1951.
7. C. Genest, B. Rémillard, D. Beaudoin. Goodness-of-fit tests for copulas: a review and a power study. *Insurance Math. Econom.*, 44, 199–213, 2009.
8. I. Gijbels, N. Veraverbeke, M. Omelka. Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis,* 55, 1919–1932, 2011.
9. W. Hoeffding. Masstabinvariante Korrelationstheorie. *Schriften des Matematischen Instituts und des Instituts für Angewandte Matematik der Universität Berlin*, 5, 179–223, 1940.
10. M. Hofert, I. Kojadinovic, M. Maechler, J. Yan. copula: Multivariate Dependence with Copulas. R package version 0.999-13, URL http://CRAN.R-project.org/package=copula , 2015.
11. H. Joe. *Dependence Modeling with Copulas*, CRC Press, Boca Raton, 2015.
12. E. Kazatchenko, M. Markov, A. Mousatov, J. Parra. Carbonate microstructure determination by inversion of acoustic and electrical data: application to a South Florida Aquifer. *J. Appl. Geophys.*, 59, 1–15, 2006.
13. D. Kurowicka. Introduction: Dependence Modeling. In D. Kurowicka and H. Joe, editors, *Dependence Modeling Vine Copula Handbook*, pp. 1–17, World Scientific Publishing, Singapore, 2011.
14. R. B. Nelsen. *An introduction to copulas*, Springer, New York, 2006.
15. B. Rémillard and J.-F. Plante. TwoCop: Nonparametric test of equality between two copulas. R package version 1.0. http://CRAN.R-project.org/package=TwoCop , 2012.
16. B. Rémillard, B. Scaillet. Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100, 377–386, 2009.
17. L. Rüschendorf. Asymptotic distributions of multivariate rank order statistics. *Ann. Statist.*, 4, 912–023, 1976.

18. K. F. Siburg and P.A. Stoimenov. Gluing copulas. *Commun. Statist.– Theory and Methods*, 37, 3124–3134, 2008.
19. A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8, 229–231, 1959.

Selecting instances with self-organizing maps: Sampling data of aircraft engines

Leonor Fernandes[1]; Roberto Henriques[2]; Victor Lobo[3]

1 NOVA IMS - NOVA Information Management School, Lisbon, Portugal
mlfernandes@euroatlantic.pt
2 NOVA IMS - NOVA Information Management School, Lisbon, Portugal
roberto@novaims.unl.pt;
3 NOVA IMS - NOVA Information Management School, Lisbon, Portugal and CINAV-PRT Navy Research Center, Almada Portugal
vlobo@novaims.unl.pt

Abstract: Aviation companies currently have a huge amount of data that enables the development of new forms of diagnosis and prognosis of faults. The application of Knowledge Discovery in Databases (KDD) and of Data Mining (DM) techniques to these data of aircraft engines allows the implementation of new Condition Based Maintenance (CBM) policies. During an flight, the amount of available data that allows determining the engine's state is abundant. However, most of the data collected are redundant, and the sheer amout of data makes all the processing very time-consuming, or even impossible with the available resources, even in a "Big Data" context. Thus, selecting significant data is an important task in the phase of pre-processing of data in KDD, and contributes to the success of the process, but it is very time-consuming. The objective of this paper is to select instances for a sample that may be used for CBM. There should be no loss of relevant information in the sample to identify the state of the engine. In this paper we have applied self-organizing maps (SOM's) to sample the data. We used the batch and the sequential algorithms to train the maps. With clustering techniques and sensibility analysis we compared the results and propose a method to choose the best sample.

Keywords: Knowledge Discovery Databases (KDD), Instance Selection, Self-Organizing Maps (SOM), non-supervised learning.

1. Introduction

To classify the state of an engine and predict the occurrence of the next failure are important information in support of the making decision to remove engines for maintenance. Usually the decision to remove the engines for repair depends on visual inspections carried out regularly, monitoring the records of the performance parameters of aircraft engines and the knowledge of experts. The monitoring of an engine's performance parameters is done by comparing the values recorded during the flight with the manufacturer's thresholds. When flight data are close to these thresholds, maintenance actions are taken. This monitoring is done individually by parameter. Due to the inherent risks of unexpected failure, engine removal is usually done before the optimal time.

Aviation companies, currently, have a register of a huge amount of data that enables the development of new forms of diagnosis and prognosis of faults. The application of Knowledge Discovery in Databases (KDD) to the registers of the engine parameters and the flight conditions, through the study of their interactions, allows another type of information about engine performance.
KDD means the process of knowledge discovery that you can obtain from the data of past experience, Gama *et al.* [1]. One of the steps in this process is data mining which consists of the application of algorithms in order to find patterns from the data.
During a flight, the data about different performance and operation parameters of the engines are recorded every second. It is possible to recognize abnormal data in order to identify possible crashes/failures in the engine, but during a flight not all data are relevant.
To select significant data is an important task in the phase of the preprocessing of data in KDD , Fernández *et al*. [2], but is time-consuming, although its tasks contribute to the success of the process Reinartz [3], mainly the step of data mining.
The objective of this is work is to use a neuronal network, SOMs, to select instances and evaluate the samples quality by using clustering techniques. We want to find a sample with the smallest dimension, in such a way that the parameters have the same behaviour as the original data.
The paper is structured as follows: Section 2 gives a brief description of the instance selection problem. The description of the data and the methodology used are in Section 3. Section 4 presents the results obtained with SOMs. Conclusions are given in Section 5.

2. Instance Selection: the problem

Currently, to use samples to make an inference about statistical population parameters is the usual procedure. The first work to present the sampling theory was at the end of sec XIX.,Seng [4]. However the proliferation of applications and data mining algorithms, over recent decades, has given rise to the need for data reduction,i.e. a reduction of data from the point of view of a decrease in the number of features (variables) and/or the number of instances (observations). In this context, in addition to the reasons which always led to the necessity to do sampling, there is also a need to select instances because the data are not recorded with the goal of applying data mining techniques or any other application. The purpose of collecting this data is operational.
Instance selection, in the KDD process, is to choose a subset of data to achieve the original purpose of the data mining, Michalski [5]. When we work with a subset of data we have gains in the steps of data mining because it is easier and faster. The sample size and quality of the results obtained in data mining are the important matters. To evaluate the best sample is a difficult task. In instances selection, the evaluation of samples depends on the objectives for which this will be used. The conventional evaluation methods for sampling, classification

and clustering can be used to assess the performance in instances selection, Liu and Motoda [6].

The literature about the instance selection is abundant. In 2001, Liu and Motoda [6] edited one book about instance selection in data mining. However, most of the works in this book are about instance selection in supervised learning. In the supervised learning we know the data of input and output space and assess the sample quality is simple, we can use per example the success rate. From this perspective, Olvera-López *et al*.[7] presents a detailed summary of the most commonly used algorithms, given their features and also comparing their performances.

The our work is about non-supervised learning. We found very little literature (after 2000) about instance selection in non-supervised learning. The works of John and Langley [8], Palmer and Faloutsos [9] ,Liu and Motoda [10]) refer to the usual sampling procedures to sample the data (i.e. simple random sampling, sampling uniform, stratified) and clustering methods. But instance selection in non-supervised learning, is important. The amount of information, in the data exploratory phase makes this analysis complex and ineffective, due to the execution time of some algorithms and the non- implementation of other algorithms Liu and Motoda [10].

In 2014, Fernandes *et al*. [11] presented a work with different methods to select instances in non-supervised learning. They used conventional methods of sampling and data mining algorithms. In their work the SOMs, with sequential algorithm showed good results but were very time-consuming.

The SOMs are non-supervised learning neural networks and use a competitive learning technique. They were originally proposed by Kohonen in 1982. Currently, SOMs are considered to be one of the most powerful learning techniques, Kohonen [12]. According Ballabio *et al.* [13] one of the major disadvantages of SOMs is to be very time-consuming due to the network optimisation.

Several parameters established before the traininig phase affect the results of the SOM. The parameters can be grouped into different types of parameter: structure parameters - size, topology, shape and initialisation of the map; and training parameters - number of iterations, learning rate, neighbourhood radius and neighbourhood function. So there are many parameters to optimize in the self-organizing maps. There are several studies about the effects of changes in the parameters of SOMs and about the compararison of the different values of those parameters (Ettaouil *et al*.[14], Cottrel *et al*. [15]).

In this work, in order to reduce the time consumed with SOMs, we used two types of algorithm to training the network: sequential and batch. The difference between these two algorithms is the way in which the data are presented to the network. In sequential data vectores are presented to the network one at a time while the batch data vectors are simultaneously presented to the network. Others changes in the parameters of SOMs were also studied.

3. Data description and Methodology

a. Data description

We have 229 flights made by one commercial B767-300 between 2009 and 2013. On each flight were recorded, second by second, the data of 25 performance parameters of the two engines and 6 features of the flight conditions. The final database is large; there are 4,232,008 instances and 31 variables. Table I describes the parameters of the database, 1/2 are the parameters of engines 1 and 2.

Table I - Description of Parameters

Parameters	
Flight Conditions (6)	Altitude
	Gross Weight
	Mach Number
	Static Air Temperature
	Total Air Temperature
	Pressure Total
Engine Performance (25)	Bleed Duct Pressure - Engine 1/2
	Temperature - Engine 1/2
	Vibration-Engine 1/2
	Pressure Ratio - Engine 1/2
	Request EPR 1/2
	Max Limit EPR1/2
	Fuel Flow - Engine 1/2
	Fan Speed-Speed Low Pressure - Engine 1/2
	Core Speed-Speed High Pressure - Engine 1/2
	Oil Pressure - Engine 1/2
	Oil Quantity - Engine 1/2
	Oil Temperature - Engine 1/2
	Throttle Resolver Angle

The number of observations per flight varies widely because it depends on the duration of each flight. Figure 1 describes the number of instances per flight. Most of flights are of more than 4 hours duration, i.e. they have a number of records greater than 15,000, Table II.

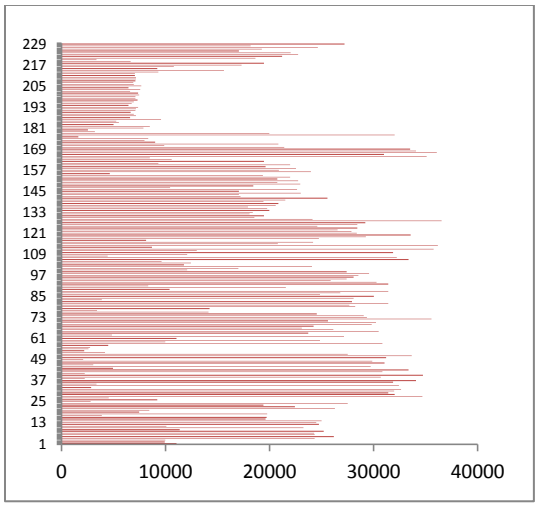

Fig. 1. Distribution of the instances per flight

Table II – Nº of flights by duration

Time of flight	Nº of flights
- 1 hour	17
1-4 hours	73
+ 4 hours	139
Total	229

b. Methodology

We have an original database which is very large so we have the typical problems when we want to apply data mining techniques. One of them is the redundancy of the values in the same flight. As the records on each flight are done every second, there are many that are equal to the previous one. The approach used in this work is done flight by flight. For the definition of the sample size we studied the behaviour of eight flights, Figure 2, selected randomly but according to the three groups of time established in Table II.

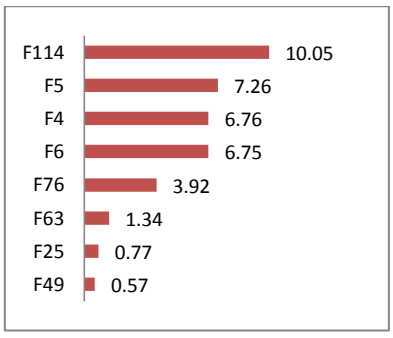

Fig. 2. Distribution of the flight time (hours)

In each flight we calculate the Euclidian distance between instance t and instance t-1.The data was standardized with z-score. We want to understand the behaviour between the distances in consecutives instances. The average of distances of each flight is less than 0.5 but have a big dispersion, Table III.

Table III – Descriptive Statistics of the Euclidian Distance between instance t and instance t-1

Statistics	Flight 4 (F4)	Flight 5 (F5)	Flight 6 (F6)	Flight 114 (F114)	Flight 49 (F49)	Flight 63 (F63)	Flight 76 (F76)	Flight 25 (F25)
Mean	0,12	0,11	0,10	0,10	0,35	0,21	0,15	0,34
Standard Deviation	0,85	0,70	0,78	0,85	1,30	0,96	0,99	1,23
Coefficient of variation	708	665	757	878	366	456	653	382
Number of records	24331	26143	24311	36187	2039	4823	14111	2779

When we analysed the differences between instances, more than 90% records have minimal distances, only 1% the records have a large distance, Figure 3.

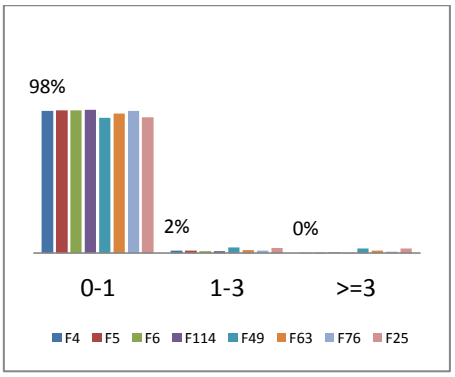

Fig. 3. Distance between instance t and t-1

So the dimension used for the sample was the 1% of the instances per flight. In this work we studied the results when some parameters changed in the SOMs. We built five scenarios which led to 11 different samples. The shape, topology initialisation of the map, learning rate, neighbourhood radius and neighbourhood function was the same for the five scenarios. The values used are in Table IV.

Table IV – SOM parameters

Structure parameters			Training parameters				
Topology	Shape	Iinitialisation	Learning rate α		Neighbourhood radius σ		Neighbourhood function
			Initial	End	Initial	End	
Hexagonal	Sheet	Randomly	0,7	0,02	8	3	Gaussian

The differences among the scenarios are because of the number of iterations, initial weights, size of maps and algorithms used to train the network.
The first four scenarios have two samples each; the difference is the training algorithm, sequential and batch. In these samples the dimension was 42.254

records. In the last scenario we used two different forms to present the data to the network in a sequential algorithm: randomly and ordered. We also used one different form to specify the size of the maps; we identified the number of units as 1% of the instances in each flight. In this scenario the size of the sample was 42.402 records. In the other´s scenarios (scenario 1 to 4) we pre-defined the size of matrices. Table V summarizes the different parameters used in each scenario.

Table V - Description of five scenarios

<table>
<tr><th rowspan="2">Scenario</th><th colspan="2">Number of iterations</th><th rowspan="2">map size</th><th rowspan="2">Initial weights (Wi)</th></tr>
<tr><th>Initial</th><th>End</th></tr>
<tr><td>1</td><td rowspan="2">20</td><td rowspan="2">30</td><td rowspan="3">229 matrices, the size of matrices varied between [4 5] and [19 20] depending flight time (almost square matrices)</td><td>Randomly but different in each algorithm</td></tr>
<tr><td>2</td><td>Randomly but equal in both algorithm</td></tr>
<tr><td>3</td><td>100</td><td>150</td><td rowspan="3">The same scenario 2</td></tr>
<tr><td>4</td><td rowspan="2">20</td><td rowspan="2">30</td><td>229 matrices, the size of matrices varied between [2 10] and [10 38] depending flight time (rectangular matrices)</td></tr>
<tr><td>5</td><td>Not defined matrices, we chose number of units=1% data on each flight</td></tr>
</table>

To assess the quality of the samples, we use one technique of clustering and the time-consuming for each scenario. In the samples and in the original database ten clusters were created for each flight. We applied the K-means algorithm for the formation of the clusters. After this, a Euclidian distance between the centroids clusters of the samples and the original database were calculated, Figure 4. We used two procedures to compare the centroids: I) According the first variable, we sorted the centroids in ascending order; II) Compared centroids of the all clusters among themselves.

At the end, we calculated the total distance by flight and compared for each procedure. The lower the total distance the greater the similarity between the sample and the original database.

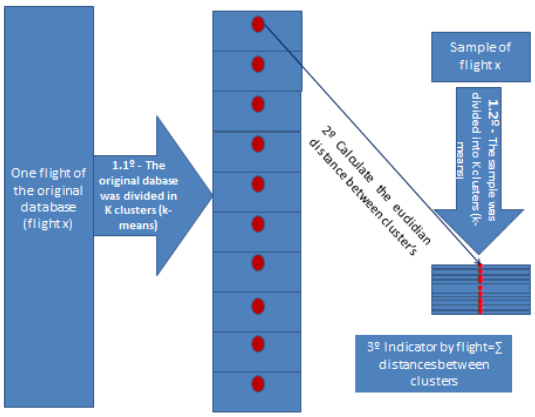

Fig. 4. Evaluation scheme of samples

At the beginning of the study we applied ten clusters because during the flight there are six natural clusters, the principal phases of flight [16], but in abnormal situations it is possible to have more phases. Later we made a sensibility analysis by using three, six and seven clusters. The six phases described in [16] are grouped usually in three: take off, cruising and landing.

4. Results

Table VI shows the total distance for the 11 samples when we used the two assess procedures and 10 clusters.

Table VI – Results for 11 samples

Scenario	Cluster's comparison			
	I - Cluster's ordered the 1st variable	II - Comparison of all cluster's	I - Cluster's ordered the 1st variable	II - Comparison of all cluster's
	SomSequential		SomBatch	
1	17.285,35	231.980,64	110.059,85	1.167.537,84
2	16.096,36	219.172,05	112.918,76	1.195.762,97
3	25.732,27	320.065,35	95.879,89	1.026.478,06
4	16.042,76	218.590,44	89.467,68	959.773,45
5.1	16.480,42	221.658,20	96.372,69	1.033.432,14
5.2	17.804,43	233.759,31		

The sequential algorithm has smaller total distances then the algorithm batch for all the scenarios. In the sequential algorithm in an orderly manner to present the data to the network not brought better distances than when we present randomly (scenarios 5.1 and 5.2). It is the way as we defined the map size that shows better results, scenario 4, where the matrix is more rectangular for both algorithms.

When we increased the number of iterations, scenario 3, the results of the sequential algorithm does not improve, unlike the success of the batch algorithm. Both algorithms present consistency in the results despite the changes in the parameters of SOMs.

The way we compare the clusters does not seem to influence the behaviour of the total distances just of magnitude values, figures bellows.

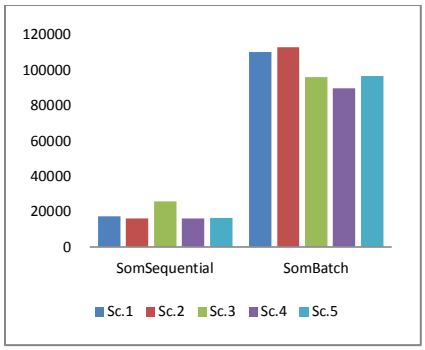

Fig. 5.- Procedure I

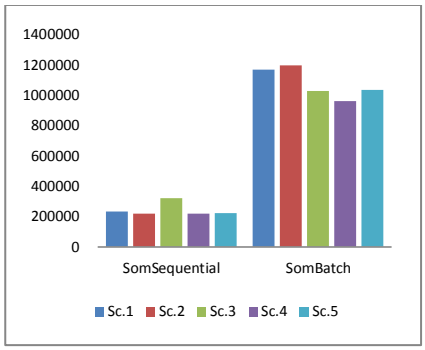

Fig. 6.- Procedure II

As regards the time-consumed, the performance of the two algorithms is the opposite: the batch algorithm consumes less time than the sequential algorithm, Table VII. While this behaviour was expected due to already published works (Cottrell *et al.*[15], Kohonen *et al.*[17]) we did not expect such a big difference.

Table VII – Time consuming in different scenarios

Time consuming (seconds)		
scenario	SomSequential	SomBatch
1	28.817,97	1.061,74
2	27.644,04	1.018,19
3	138.224	4.880,60
4	28.378,00	1.045,80
5.1	36.176,11	1.210,69
5.2	36.401,01	

The previous results of the algorithm batch, time consuming and better results obtained with an increase in the number of iterations led us to create more four scenarios increasing the number of iterations. These new simulations had scenario 5 as their base. If this new simulations obtain better final results, this will compensate for the increase in time-consuming when compared with the sequential algorithm.

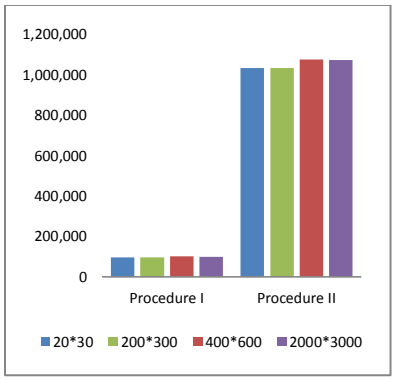

Fig. 7.– Total distance for different number of iterations

Table VIII – Time-consuming when increasing the number of iterations

Nº of iterations	Time (seconds)
2000*3000	120.720,99
200*300	11.738,16
400*600	21.778,10
20*30	1.210,69

The increase in the number of iterations did not change the results, Figure 7, and significantly increase the time-consuming in the training of SOMs, Table VIII. As the two procedures used to compare the clusters led to the same conclusion we undertook sensibility analyses with the number of clusters, Table IX. Here we also used scenario 5 as the base. The shaded values refer to the situation in which the data were presented in an orderly way to the network.

Table IX – Sensibility analyses of number of clusters

Nº of clusters	Cluster's comparison			
	I - Cluster's ordered the 1st variable	II - Comparison of all cluster's	I - Cluster's ordered the 1st variable	II - Comparison of all cluster's
	SomSequential		SomBatch	
K=3	2.380,58	15.473,92	12.427,43	45.967,63
	2.519,89	15.977,75		
K=6	8.168,54	73.755,05	54.536,37	355.111,03
	8.278,79	73.652,51		
K=7	10.043,69	100.995,69	62.230,72	470.794,69
	11.102,29	108.781,46		
K=10	16.480,42	221.658,20	96.372,69	1.033.432,14
	17.804,43	233.759,31		

The total distances increase with the number of clusters in both algorithms and evaluation procedures, Figure 8 and 9. There are no differences when the data are presented to the network in an orderly or random way.

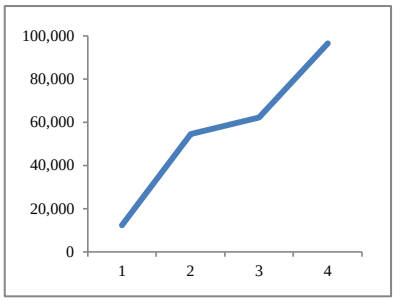

Fig. 8.– Procedure I-SomBatch

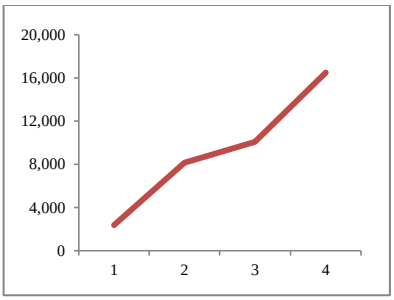

Fig. 9.- Procedure I – SomSequential

5. Conclusion

Several simulations were carried out with the objective of constituting a representative sample of the original data with dimension of approximately 1% of the original.

The optimization of SOM depends on the values of the parameters. The results for the different scenarios were consistent. The big difference was in the algorithm used. The sequential algorithm has always had better results than the batch algorithm. Even when increasing the number of iterations within the batch algorithm the values obtained are worse than those presented by the sequential algorithm and the time- consuming greatly increases. The disadvantage of the sequential algorithm is the time-consuming in the train of the network; in this case it is far superior to the batch. Even when presenting the data to the network in an orderly manner there are no gains.

As sampling the data is not an everyday procedure in the process of KDD, i.e. the same sample is used for applying different data mining techniques; it seems to us to be more relevant that the results are obtained in terms of similarity to the original data instead of the regarding the time consumed, so we preferred the sequential algorithm.

In the evaluation of the results of each sample the way to compare the clusters did not show relevant difference but the numbers of clusters used led to variations of the total distances.

References

[1] Gama J.; Carvalho A.;Faceli K.; Lorena A.; Oliveira M., *Extração de Conhecimento de Dados* 1ª edição ed. Lisboa, 2012.

[2] Fernández A.; Duarte A.; Hernández R.; Sánchez Á., "GRASP for Instance Selection in Medical Data Sets," in *Advances in Bioinformatics*. vol. 74, M. Rocha, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 53-60.

[3] Reinartz T., "A Unifying View on Instance Selection," *Data Mining and Knowledge Discovery*, vol. 6, pp. 191-210, 2002/04/01 2002.

[4] Seng Y. P., "Historical Survey of the Development of Sampling Theories and Practice," *Journal of the Royal Statistical Society. Series A (General)*, vol. 114, pp. 214-231, 1951.

[5] Michalski R.S., "On the Selection of Representative Samples from Large Relational Tables for Inductive Inference," Department of Engineering, University of Illinois at Chicago Circle, Chicago1975.

[6] Liu H.; Motoda H., *Instance Selection and Construction for Data Mining*, 1 ed.: Springer US, 2001.

[7] Olvera-López J. A.; Carrasco-Ochoa J. A.; Martínez-Trinidad J. F.; Kittler J., "A review of instance selection methods," *Artificial Intelligence Review*, vol. 34, pp. 133-143, 2010/08/01 2010.

[8] John G.; Langley P., "Static Versus Dynamic Sampling for Data Mining," in *Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, 1996, pp. 367-370.

[9] Palmer R.C.; Faloutsos C., "Density biased sampling: an improved method for data mining and clustering," *SIGMOD Rec.*, vol. 29, pp. 82-92, 2000.

[10] Liu H.; Motoda H., "On Issues of Instance Selection," *Data Mining and Knowledge Discovery*, vol. 6, pp. 115-130, 2002/04/01 2002.

[11] Fernandes L.; Henriques R.; Lobo V., "Selection of instances in Condition Based Monitoring: the case of aircraft engines.," in *Maintenance Performance Measurement and Management (MPMM) Conference 2014*, Coimbra, 2014, p. 228.

[12] Kohonen T., *Self-Organization and Associative Memory* vol. 8: Springer Berlin Heidelberg, 1989.

[13] D. Ballabio, *et al.*, "Effects of supervised Self Organising Maps parameters on classification performance," *Analytica Chimica Acta*, vol. 765, pp. 45-53, 2013.

[14] Ettaouil M.; Abdelatifi E.; Belhabib F.; Moutaouakil K. E., "Learning Algorithm of Kohonen Network With Selection Phase," *WSEAS Transactions on Computers*, vol. 11, p. 387, Novembro 2012.

[15] Cottrel M.; Fort J.C. ; Letremy P., "Advantages and drawbacks of the batch Kohonen algorithm," presented at the 10th European Symp. On Artificial Neural Network, Bruges, Belgium, 2005.

[16] C. Common Taxonomy Team. (2012) Phase Of Flight - Definitions And Usage Notes.

[17] Kohonen T.; Kaski S.;Lagus K.;Salojarvi J.;Honkela J.;Paatero V.;Saarela A., "Self organization of a massive document collection," *Neural Networks, IEEE Transactions on*, vol. 11, pp. 574-585, 2000.

Computations of retirement age based on generation life tables

Tomas Fiala[1] and Jitka Langhamrova[2]

[1] Department of Demography, Faculty of Informatics and Statistics, University of Economics, Prague, nam. W. Churchilla 4, 130 67 Praha 3, Czech Republic
(E-mail: fiala@vse.cz)
[2] Department of Demography, Faculty of Informatics and Statistics, University of Economics, Prague, nam. W. Churchilla 4, 130 67 Praha 3, Czech Republic
(E-mail: langhamj@vse.cz)

Abstract. According to current legislation the statutory retirement age is planned to increase permanently with constant increment for each subsequent generation. The Council of the European Union recommends to link the rise more clearly with expected changes in life expectancy.
The Expert Committee on Pension Reform of the Czech Republic recommends to determine the retirement age as the age when the percentage share of the life expectancy (calculated as the arithmetic mean of these expectancies for males and females) in respect of the total expected length of life at that age will be about 25%, i.e. one quarter of their total expected length of life. Of course the generation life tables should be used for this computations.
The paper presents computation of the retirement age in the case of the Czech Republic according to generation life tables based on the mortality scenarios of the latest population projection by the Czech Statistical Office and on the latest Eurostat population projection. In both projections the retirement age in the second half of this century should be lower than the present legislation values.
Keywords: Generation life tables, pension system, retirement age.

1 Introduction

One of the themes dealt with in 2014 by the Expert Committee on Pension Reform of the Czech Republic was the adjustment of the age limit for retirement in the Czech Republic, attaining which is one of the conditions for entitlement to the old-age pension. According to the existing legal arrangement (Law No. 155/1995 Coll., Zákon [9]) the retirement age is to permanently rise in time linearly regardless of the development of the life expectancy in the Czech Republic. Theoretically, therefore, it could happen that, if the growth of life expectancy slowed down or stopped, in the future many people would not even reach retirement age or would receive a pension only for a relatively short period of time.
Completely halting any further rise in retirement age after reaching a certain limit (such as 65 years), as proposed by some political parties, would not,

however, be suitable. The Council of the European Union, on the contrary, recommends the Czech Republic to *Ensure the long-term sustainability of the public pension scheme, in particular by accelerating the increase of the statutory retirement age and then by linking it more clearly to changes in life expectancy.* (Council Recommendation [2], p. 15). According to the present projection of the population of the Czech Republic (CZSO [3]) and other prognoses (e.g. Burcin and Kučera [1]), it is expected that in the Czech Republic there will be a permanent rise in the life expectancy throughout the present century. The continuing rise of the retirement age is thus in accordance with the recommendation of the European Council.

The question remains of how to link the retirement age with the development of the life expectancy to achieve a certain degree of stabilisation of the average period of receipt of the old-age pension. Proposals have appeared, for example, which envisaged selecting the retirement age in such a way that the average duration of the receipt of the old-age pension would be roughly constant, say 20 years. This would mean, however, that, assuming a rise in the length of life, the retirement age would rise and thus also the expected length of economic activity, but the period of receipt of the pension would remain the same and the relative period of receipt of the pension would drop.

The Expert Committee on Pension Reform therefore finally approved the recommendation that the retirement age (which should be the same for both men and women) and, as hitherto, should depend on the year of birth of the individual. The value of the retirement age should in this case be determined so that people reaching senior age should receive an old-age pension on average for the last quarter of their lives (Expert Committee [5]). For the generations born before 1966 the retirement age in this case should continue to be in accordance with the present legislation, where for every generation of succeeding year of birth the retirement age rises (in comparison with the preceding year generation) for men by 2 months. For women (which have at present time lower retirement age) the rise is by 4 months, after 2019 by 6 months until they reach the level of men. Men born in 1965 should thus retire at the age of 65 years, women with 2 children at 64 years 8 months.

The calculations for the average period of receipt of old-age pensions must, however, be based on the life expectancy at the appropriate age (not at birth) and also calculated on the basis of the generation (not cross-section) life tables. The cross-section life tables, which are usually published for each year, give the life expectancy on the assumption that mortality in the years to come will still be the same as in the year of the tables. To estimate the mortality and life expectancy of the actual population it is therefore necessary to use the generation mortality tables, which take into account the expected changes in mortality in later years (e.g. Roubíček, [8]).

The aim of this paper is to present model calculations of the development of retirement age in the Czech Republic with the application of the above-mentioned proposal that seniors should receive their pension on average for the last quarter of their lives. The calculations are made for generations born in the period 1950–2020, i.e. for generations which will reach the retirement age in the

period since the present time until almost the end of this century. The calculation of the generation life expectancies and hence also the values for retirement age was carried out in two variants of the expected development of mortality in the Czech Republic. The first is based on the scenario of the mortality of the medium variant of the projection of the Czech population 2013 (CZSO [3]), the second is based on the main scenario of the Eurostat projection for the Czech Republic from the same year (Eurostat [4]).
Throughout this paper pension will be taken to mean the lifelong old-age pension paid out from the moment of attaining retirement age until the deaths (unless expressly stated otherwise).

2 Methodological notes

The most accurate estimate of the real length of the remaining life of a person at age x of birth year g is the life expectancy $e_x^{(g)}$ of a person at the age x from the generation mortality tables for the generation of birth year g. Because the mortality of males and females differs, usually the mortality tables and thus also the life expectancy are calculated separately by gender. For the calculation of retirement age, which is to be the same for both men and women, we will use the life expectancy without gender differentiation defined for each age unit as the arithmetical mean of the life expectancies of males and females at this age.
The estimate of the average length of entire life of a person of birth year g at the moment they attain the age of x, is understandably then the value $x+e_x^{(g)}$. According to the proposal of the Pension Committee the theoretical retirement age $x^{(g)}$ for the generation of those born in year g should be such that for this generation it applies that

$$\frac{e_x}{x+e_x} = 0.25 \qquad (1).$$

(upper index g is omitted for simplicity). The expected average period of receipt of pension for this generation $e_x^{(g)}$ would then be equal to a third of the value of their retirement age $x^{(g)}$.
Life expectancies are usually calculated only for integer age values. When seeking the value of retirement age we therefore find from the mortality tables for the appropriate generation first of all the highest integer age value x, for which the above-mentioned share (1) is still higher or equal to 0.25; then we determine the "more accurate" value of retirement age by linear interpolation between values x and $x+1$ where

$$\frac{e_x}{x+e_x} \geq 0.25 \quad \text{and} \quad \frac{e_{x+1}}{x+1+e_{x+1}} < 0.25 \qquad (2).$$

Because life expectancy declines with the increase in age, the solution is always unequivocal.
For determining the retirement age of the generation born in year g it is not, therefore, necessary to have the values of life expectancies for the entire age range. It is sufficient to know these values for the higher age when the life expectancy comes close to a quarter of the average length of entire life. Present

mortality prognoses indicate that for the generations born after 1950 this age will be over 60 years for both males and females.

Calculation of generation life expectancies

The life expectancy of a person at the age of x is influenced only by mortality at this age and higher ages and does not depend on mortality at lower ages. For the calculation of life expectancies of the generation of persons born in year g at the age of 60 and over it therefore suffices to know for each generation g the age-specific mortality rates $m_x^{(g)}$ of this generation for the age $x \geq 60$. We then calculate the life expectancies in the following manner (upper index g designating the generation is omitted for simplicity):

We select the initial value for the number survivors l_{60}, for $x \geq 60$ we calculate the survival probability and the number of survivors to the age $x+1$

$$p_x = e^{-m_x}, \qquad l_{x+1} = l_x \cdot p_x . \tag{3}$$

The total number of person-years lived above the age x is then

$$T_x = \sum_{u=x}^{\omega-1} L_u = \frac{l_x + l_{x+1}}{2} + \frac{l_{x+1} + l_{x+2}}{2} + \ldots + \frac{l_{\omega-1} + l_\omega}{2} = \sum_{u=x}^{\omega-1} l_u - \frac{l_x}{2}, \tag{4}$$

and the life expectancy at the age x

$$e_x = \frac{T_x}{l_x} = \frac{\sum_{u=x}^{\omega-1} l_u}{l_x} - \frac{1}{2}, \tag{5}$$

its value does not depend on the selected initial value of the number of survivors l_{60}.

Estimate of generation age-specific mortality rates by cross-section rates.

The values of generation age-specific mortality rates can be acquired in several ways. One of these is an estimate based on (real or forecasted) cross-section mortality rates. A person born in the year g will reach the age of x in the year $g+x$, and will therefore live at the completed age x partly in the year $g+x$, partly in the year $g+x+1$. As an estimate of the mortality rate at age x of the generation born in year g we can use the mean of the cross-section mortality rates

$$m_x^{(g)} = \frac{m_{g+x,x} + m_{g+x+1,x}}{2}, \tag{6}$$

where $m_{t,x}$ is the cross-section mortality rate in year t at the age x.

Mortality scenarios of the population projections of the Czech Republic according to the Czech Statistical Office – CZSO (medium variant) and Eurostat (main variant).

The mortality scenarios of both projections are based on the expectation of the continuing reduction in the mortality of both males and females and therefore consider the growth of the life expectancy at birth for both genders during the entire period of the projection. There is a difference, however, in the rate of this growth. The projection of the CZSO envisages more rapid growth of the life expectancy up to 2030, but after that (up to 2100) a gradual slowing of the

annual rise in the life expectancy, whereas Eurostat starts from the assumption of a lower, but more stable, rise in life expectancy for the whole period up to 2080. The CZSO projection thus considers a life expectancy for both males and females in the middle of the century roughly 1 year 4 months higher than Eurostat, but gradually the difference between the two projections drops and roughly from the seventies Eurostat expects a higher life expectancy than the CZSO.

CZSO did not publish the values of age-specific mortality rates for the individual years in the scenario of its projections. An estimate of these was therefore made by the projected proportions of the number of persons living. For the age of 60 years and over it is possible to assume that foreign migration is negligible and so the survival probability for age of x in year t was estimated according to the formula

$$p_{t,x} = \frac{S_{t+1,x} + S_{t+1,x+1}}{S_{t,x-1} + S_{t,x}} . \qquad (7)$$

Eurostat directly states the values of the specific mortality levels used for the calculation of the projection.

For the calculation of life expectancies for generations of those born in the years 1950–2020 according to the formula (6) the values of cross-section age-specific mortality rates at the age from 60 years and more are needed for each year of the period 2010–2125 (we assume that nobody will reach 105 years). The estimate of specific mortality rates, or survival probabilities for further years (from the horizon of the projection up to 2125) will then be carried out on the assumption that the annual rate of the drop in mortality for individual units of age in the following years will be the same as in the last decade of the projection scenario.

The generation age-specific mortality rates or survival probabilities were then calculated analogically as (6).

Comparison of the generation life expectancies of both projection scenarios

For the calculation of retirement age, as has already been said, the life expectancy at birth is not important, but first and foremost the life expectancies of the individual generations at the age of 60 and over. These values for selected units of age, where it may be expected that people will have roughly the last quarter of their lives before them, are given in Tab. 1.

From the difference in the two scenarios described above it is evident that the projection of the CZSO envisages a slightly higher life expectancy for the older generations and a slightly lower life expectancy for younger generations in comparison with the Eurostat projection. The borderline is formed by the generations born around 1985, for which the life expectancy is roughly the same according to both projections. The life expectancies of women are roughly 3–5 years higher than the life expectancies of men, but the differences lessen with age.

Table 1. Generation life expectancies for selected ages
(mean of values for males and females)

Year of births	Czech Statistical Office Scenario (medium variant)											Eurostat Main Scenario									
	age											age									
	60	61	62	63	64	65	66	67	68	69	70	60	61	62	63	64	65	66	67	68	69	70
1950	24,07	23,32	22,57	21,84	21,11	20,38	19,66	18,95	18,24	17,54	16,84	23,39	22,63	21,88	21,13	20,39	19,65	18,92	18,20	17,48	16,77	16,07
1951	24,27	23,51	22,76	22,02	21,28	20,55	19,83	19,11	18,39	17,69	16,98	23,56	22,79	22,03	21,28	20,53	19,78	19,05	18,32	17,60	16,89	16,18
1952	24,45	23,69	22,94	22,19	21,45	20,71	19,98	19,26	18,54	17,83	17,12	23,72	22,95	22,18	21,41	20,66	19,91	19,17	18,44	17,72	17,00	16,29
1953	24,64	23,87	23,11	22,36	21,62	20,88	20,14	19,41	18,69	17,97	17,25	23,87	23,09	22,32	21,55	20,79	20,04	19,30	18,56	17,84	17,11	16,40
1954	24,82	24,05	23,29	22,53	21,78	21,04	20,30	19,56	18,83	18,10	17,38	24,02	23,24	22,46	21,68	20,92	20,17	19,42	18,68	17,95	17,23	16,51
1955	24,99	24,22	23,45	22,69	21,94	21,19	20,44	19,70	18,97	18,23	17,51	24,17	23,38	22,59	21,82	21,05	20,30	19,55	18,80	18,07	17,34	16,62
1956	25,16	24,38	23,61	22,85	22,09	21,33	20,58	19,84	19,10	18,36	17,62	24,31	23,52	22,73	21,95	21,18	20,42	19,67	18,92	18,19	17,45	16,73
1957	25,32	24,54	23,77	23,00	22,23	21,47	20,72	19,97	19,22	18,47	17,73	24,45	23,65	22,87	22,09	21,31	20,55	19,79	19,04	18,30	17,57	16,84
1958	25,48	24,70	23,92	23,15	22,38	21,61	20,85	20,09	19,34	18,58	17,84	24,59	23,79	23,00	22,22	21,44	20,67	19,91	19,16	18,42	17,68	16,94
1959	25,66	24,87	24,09	23,31	22,53	21,76	20,99	20,23	19,46	18,71	17,95	24,73	23,93	23,14	22,35	21,57	20,80	20,04	19,28	18,53	17,79	17,05
1960	25,83	25,03	24,24	23,46	22,68	21,90	21,13	20,36	19,59	18,82	18,06	24,87	24,07	23,27	22,48	21,70	20,92	20,16	19,40	18,65	17,90	17,16
1961	25,98	25,18	24,39	23,60	22,81	22,03	21,25	20,47	19,70	18,92	18,16	25,01	24,20	23,40	22,61	21,83	21,05	20,28	19,52	18,76	18,01	17,27
1962	26,13	25,32	24,53	23,73	22,94	22,16	21,37	20,59	19,81	19,03	18,26	25,15	24,34	23,54	22,74	21,95	21,17	20,40	19,63	18,87	18,12	17,37
1963	26,26	25,46	24,66	23,86	23,06	22,27	21,48	20,69	19,90	19,12	18,35	25,29	24,48	23,67	22,87	22,08	21,30	20,52	19,75	18,99	18,23	17,48
1964	26,38	25,57	24,76	23,96	23,16	22,36	21,57	20,77	19,98	19,20	18,43	25,43	24,61	23,80	23,00	22,20	21,42	20,64	19,87	19,10	18,34	17,59
1965	26,50	25,69	24,88	24,07	23,26	22,46	21,66	20,86	20,07	19,29	18,51	25,56	24,74	23,93	23,13	22,33	21,54	20,76	19,98	19,21	18,45	17,69
1966	26,63	25,81	25,00	24,18	23,37	22,56	21,76	20,96	20,17	19,38	18,60	25,70	24,88	24,06	23,25	22,45	21,66	20,87	20,10	19,32	18,56	17,80
1967	26,75	25,93	25,11	24,29	23,48	22,66	21,86	21,05	20,26	19,47	18,68	25,83	25,01	24,19	23,38	22,58	21,78	20,99	20,21	19,43	18,66	17,90
1968	26,87	26,04	25,22	24,40	23,58	22,76	21,95	21,15	20,35	19,56	18,77	25,97	25,14	24,32	23,51	22,70	21,90	21,11	20,32	19,54	18,77	18,00
1969	26,98	26,15	25,32	24,49	23,67	22,85	22,04	21,24	20,43	19,64	18,85	26,10	25,27	24,45	23,63	22,82	22,02	21,22	20,44	19,65	18,88	18,11
1970	27,07	26,24	25,41	24,58	23,76	22,94	22,12	21,31	20,51	19,71	18,92	26,23	25,40	24,57	23,76	22,94	22,14	21,34	20,55	19,76	18,98	18,21
1971	27,16	26,33	25,50	24,67	23,84	23,02	22,20	21,39	20,58	19,78	18,99	26,36	25,53	24,70	23,88	23,06	22,26	21,46	20,66	19,87	19,09	18,31
1972	27,26	26,42	25,58	24,75	23,93	23,10	22,29	21,47	20,66	19,86	19,06	26,49	25,66	24,83	24,00	23,19	22,37	21,57	20,77	19,98	19,19	18,41
1973	27,34	26,50	25,67	24,83	24,00	23,18	22,36	21,54	20,73	19,92	19,12	26,62	25,79	24,95	24,13	23,30	22,49	21,68	20,88	20,09	19,30	18,52
1974	27,41	26,57	25,73	24,90	24,07	23,24	22,42	21,60	20,78	19,97	19,17	26,75	25,91	25,08	24,25	23,42	22,61	21,80	20,99	20,20	19,40	18,62
1975	27,49	26,64	25,80	24,97	24,13	23,30	22,48	21,65	20,84	20,02	19,21	26,88	26,04	25,20	24,37	23,54	22,72	21,91	21,10	20,30	19,51	18,72
1976	27,57	26,72	25,88	25,04	24,20	23,37	22,54	21,71	20,89	20,08	19,26	27,01	26,16	25,32	24,49	23,66	22,84	22,02	21,21	20,41	19,61	18,82
1977	27,65	26,80	25,95	25,11	24,27	23,44	22,60	21,77	20,95	20,13	19,31	27,14	26,29	25,44	24,61	23,78	22,95	22,13	21,32	20,51	19,71	18,92
1978	27,72	26,87	26,02	25,18	24,34	23,50	22,66	21,83	21,00	20,18	19,36	27,26	26,41	25,57	24,73	23,89	23,06	22,24	21,43	20,62	19,81	19,02
1979	27,79	26,94	26,09	25,24	24,39	23,55	22,71	21,88	21,05	20,22	19,39	27,39	26,53	25,69	24,84	24,01	23,18	22,35	21,53	20,72	19,92	19,11
1980	27,86	27,01	26,15	25,30	24,45	23,61	22,76	21,93	21,09	20,26	19,43	27,51	26,66	25,81	24,96	24,12	23,29	22,46	21,64	20,83	20,02	19,21
1981	27,94	27,08	26,23	25,37	24,52	23,67	22,83	21,98	21,15	20,31	19,48	27,64	26,78	25,92	25,08	24,24	23,40	22,57	21,75	20,93	20,12	19,31
1982	28,01	27,15	26,29	25,43	24,58	23,73	22,88	22,03	21,19	20,35	19,52	27,76	26,90	26,04	25,19	24,35	23,51	22,68	21,85	21,03	20,22	19,41
1983	28,08	27,21	26,35	25,49	24,63	23,77	22,92	22,07	21,23	20,39	19,56	27,88	27,02	26,16	25,31	24,46	23,62	22,79	21,96	21,13	20,32	19,50
1984	28,14	27,27	26,40	25,54	24,68	23,82	22,97	22,12	21,27	20,43	19,60	28,00	27,14	26,28	25,42	24,57	23,73	22,89	22,06	21,24	20,41	19,60
1985	28,19	27,32	26,45	25,59	24,72	23,86	23,01	22,15	21,31	20,47	19,64	28,12	27,25	26,39	25,54	24,68	23,84	23,00	22,16	21,34	20,51	19,69
1986	28,25	27,37	26,50	25,63	24,77	23,90	23,05	22,19	21,35	20,51	19,67	28,24	27,37	26,51	25,65	24,79	23,95	23,10	22,27	21,44	20,61	19,79
1987	28,30	27,43	26,55	25,68	24,81	23,95	23,09	22,24	21,39	20,55	19,71	28,36	27,49	26,62	25,76	24,90	24,05	23,21	22,37	21,54	20,71	19,88
1988	28,35	27,47	26,60	25,72	24,85	23,99	23,13	22,28	21,43	20,59	19,75	28,48	27,60	26,73	25,87	25,01	24,16	23,31	22,47	21,63	20,80	19,98
1989	28,40	27,52	26,64	25,77	24,90	24,03	23,17	22,32	21,47	20,62	19,79	28,59	27,72	26,85	25,98	25,12	24,27	23,42	22,57	21,73	20,90	20,07
1990	28,45	27,57	26,69	25,81	24,94	24,08	23,21	22,36	21,51	20,66	19,83	28,71	27,83	26,96	26,09	25,23	24,37	23,52	22,67	21,83	20,99	20,16
1991	28,49	27,61	26,73	25,85	24,98	24,11	23,25	22,39	21,54	20,70	19,86	28,82	27,95	27,07	26,20	25,34	24,48	23,62	22,77	21,93	21,09	20,25
1992	28,54	27,66	26,77	25,90	25,02	24,16	23,29	22,43	21,58	20,74	19,90	28,94	28,06	27,18	26,31	25,44	24,58	23,72	22,87	22,02	21,18	20,35
1993	28,59	27,70	26,82	25,94	25,06	24,19	23,33	22,47	21,62	20,77	19,93	29,05	28,17	27,29	26,42	25,55	24,68	23,82	22,97	22,12	21,28	20,44
1994	28,63	27,74	26,86	25,98	25,10	24,23	23,36	22,50	21,65	20,80	19,96	29,16	28,28	27,40	26,52	25,65	24,78	23,92	23,07	22,22	21,37	20,53
1995	28,69	27,80	26,91	26,03	25,15	24,28	23,41	22,55	21,70	20,85	20,01	29,28	28,39	27,51	26,63	25,75	24,89	24,02	23,16	22,31	21,46	20,62
1996	28,75	27,86	26,97	26,09	25,21	24,33	23,47	22,60	21,75	20,90	20,05	29,39	28,50	27,61	26,73	25,86	24,99	24,12	23,26	22,40	21,55	20,71
1997	28,81	27,91	27,02	26,14	25,26	24,38	23,51	22,65	21,79	20,94	20,10	29,50	28,61	27,72	26,84	25,96	25,09	24,22	23,36	22,50	21,64	20,79
1998	28,85	27,96	27,07	26,18	25,30	24,42	23,55	22,69	21,83	20,98	20,13	29,61	28,71	27,83	26,94	26,06	25,19	24,32	23,45	22,59	21,73	20,88
1999	28,90	28,00	27,11	26,22	25,34	24,46	23,59	22,73	21,87	21,02	20,17	29,71	28,82	27,93	27,04	26,16	25,28	24,41	23,55	22,68	21,82	20,97
2000	28,94	28,05	27,15	26,27	25,38	24,51	23,63	22,77	21,91	21,06	20,21	29,82	28,93	28,03	27,15	26,26	25,38	24,51	23,64	22,77	21,91	21,06
2001	28,99	28,09	27,20	26,31	25,42	24,55	23,67	22,81	21,95	21,09	20,25	29,93	29,03	28,14	27,25	26,36	25,48	24,60	23,73	22,87	22,00	21,14
2002	29,03	28,13	27,24	26,35	25,47	24,59	23,71	22,85	21,98	21,13	20,28	30,03	29,14	28,24	27,35	26,46	25,58	24,70	23,83	22,96	22,09	21,23
2003	29,07	28,18	27,28	26,39	25,51	24,63	23,75	22,89	22,02	21,17	20,32	30,14	29,24	28,34	27,45	26,56	25,67	24,79	23,92	23,05	22,18	21,32
2004	29,12	28,22	27,33	26,44	25,55	24,67	23,79	22,93	22,06	21,21	20,36	30,24	29,34	28,44	27,55	26,65	25,77	24,89	24,01	23,13	22,27	21,40
2005	29,16	28,26	27,37	26,48	25,59	24,71	23,83	22,96	22,10	21,25	20,40	30,35	29,44	28,54	27,64	26,75	25,86	24,98	24,10	23,22	22,35	21,48
2006	29,20	28,30	27,41	26,52	25,63	24,75	23,87	23,00	22,14	21,28	20,43	30,45	29,54	28,64	27,74	26,85	25,96	25,07	24,19	23,31	22,44	21,57
2007	29,24	28,34	27,45	26,55	25,67	24,79	23,91	23,04	22,18	21,32	20,47	30,55	29,64	28,74	27,84	26,94	26,05	25,16	24,28	23,40	22,52	21,65
2008	29,27	28,37	27,48	26,59	25,70	24,82	23,94	23,07	22,21	21,35	20,50	30,65	29,74	28,84	27,93	27,04	26,14	25,25	24,37	23,48	22,61	21,73
2009	29,31	28,41	27,52	26,63	25,74	24,86	23,98	23,11	22,24	21,38	20,53	30,75	29,84	28,93	28,03	27,13	26,23	25,34	24,45	23,57	22,69	21,82
2010	29,36	28,46	27,56	26,67	25,78	24,89	24,01	23,14	22,28	21,42	20,57	30,85	29,94	29,03	28,12	27,22	26,32	25,43	24,54	23,66	22,77	21,90
2011	29,41	28,51	27,61	26,71	25,82	24,94	24,06	23,18	22,32	21,46	20,60	30,95	30,04	29,13	28,22	27,31	26,41	25,52	24,63	23,74	22,86	21,98
2012	29,46	28,56	27,66	26,76	25,87	24,98	24,10	23,23	22,36	21,50	20,65	31,05	30,13	29,22	28,31	27,41	26,50	25,61	24,71	23,83	22,94	22,06
2013	29,51	28,60	27,70	26,80	25,91	25,02	24,14	23,27	22,40	21,54	20,68	31,14	30,23	29,31	28,40	27,50	26,59	25,69	24,80	23,91	23,02	22,14
2014	29,55	28,65	27,75	26,85	25,95	25,07	24,18	23,31	22,44	21,58	20,72	31,24	30,32	29,41	28,49	27,59	26,68	25,78	24,88	23,99	23,10	22,22
2015	29,60	28,69	27,79	26,89	26,00	25,11	24,22	23,35	22,48	21,61	20,76	31,33	30,41	29,50	28,59	27,68	26,77	25,87	24,97	24,07	23,18	22,30
2016	29,65	28,74	27,84	26,94	26,04	25,15	24,27	23,39	22,52	21,65	20,79	31,43	30,51	29,59	28,68	27,76	26,86	25,95	25,05	24,16	23,26	22,37
2017	29,70	28,79	27,88	26,98	26,08	25,19	24,31	23,43	22,56	21,69	20,83	31,52	30,60	29,68	28,76	27,85	26,94	26,04	25,13	24,24	23,34	22,45
2018	29,74	28,83	27,93	27,02	26,13	25,23	24,35	23,47	22,59	21,73	20,87	31,61	30,69	29,77	28,85	27,94	27,03	26,12	25,22	24,32	23,42	22,53
2019	29,79	28,88	27,97	27,07	26,17	25,27	24,39	23,51	22,63	21,77	20,90	31,71	30,78	29,86	28,94	28,02	27,11	26,20	25,30	24,40	23,50	22,60
2020	29,84	28,92	28,02	27,11	26,21	25,32	24,43	23,55	22,67	21,80	20,94	31,80	30,87	29,95	29,03	28,11	27,20	26,29	25,38	24,48	23,58	22,68

Source: Author's computations

Model retirement age ensuring the receipt of a pension on average for the last quarter of one's life

The retirement age for individual generations, assuming that the average period of receipt of a pension should equal a quarter of the average lengths of entire life of seniors, is given in Tab. 2. For greater clarity the values of retirement age are rounded up into whole months. From the differences in the generation life expectancies in retirement age of the two scenarios it emerges that according to the CZSO scenario the retirement age for generations born roughly up to 1985 would be slightly higher than according to the Eurostat scenario and for generations born later it should be the other way round.

In both scenarios it is evident, however, that the proposal of the Pension Committee, even if it were realised immediately, decidedly would not lead to a reduction in the retirement age of the generations entering retirement at present. The retirement age according to present legislation is, for generations born in the fifties of last century, roughly one year for men and sometimes several years for women (according to the number of children reared) lower than the model value envisaging the average period of receipt of a pension as the last quarter of one's life.

For the model calculations according to the CZSO projection the retirement age for men born in 1965 according to present legislation is still 8 months lower than the model value. The model value for retirement age according to the Eurostat projection is, however, already 1 month lower for this generation than the 65 years hitherto proposed.

For younger generations, however, it is evident that with the development of mortality according the assumptions of the Eurostat scenario the retirement age for further generations would be higher according to present legislation than the retirement age ensuring that the average duration of receipt of a pension would equal the length of a quarter of a lifetime and that this difference would increase. For the generation of those born in 2000 this difference would already be roughly 2 ¾ years and for those born in 2020 the difference would be almost 5 years.

With the development of mortality according to the CZSO projection the retirement age according to present legislation would always be lower for the generations born up to 1973 than the model retirement age, but then, with regard to the expected strong slowing of the growth of life expectancy and the continuing raising of the retirement age by 2 months for each generation, the difference would increase rapidly. For the generation of those born in 2020 the retirement age would, according to present legislation, already be more than 6 years higher than the retirement age ensuring receipt of a pension for the last quarter of one's life on average.

Table 2. Current and proposed retirement age in the Czech Republic

Year of births	Receipt of pension thle last quarter of life		Current legislation	
	CZSO scenario	Eurostat scenario	males	females 2 children[1]
1950	63 9/12	63 2/12	62 6/12	58 4/12
1951	63 11/12	63 3/12	62 8/12	58 8/12
1952	64 1/12	63 5/12	62 10/12	59
1953	64 3/12	63 6/12	63	59 4/12
1954	64 5/12	63 8/12	63 2/12	59 8/12
1955	64 7/12	63 9/12	63 4/12	60
1956	64 8/12	63 10/12	63 6/12	60 4/12
1957	64 10/12	64	63 8/12	60 8/12
1958	64 11/12	64 1/12	63 10/12	61 2/12
1959	65 1/12	64 3/12	64	61 8/12
1960	65 3/12	64 4/12	64 2/12	62 2/12
1961	65 4/12	64 5/12	64 4/12	62 8/12
1962	65 5/12	64 7/12	64 6/12	63 2/12
1963	65 6/12	64 8/12	64 8/12	63 8/12
1964	65 7/12	64 9/12	64 10/12	64 2/12
1965	65 8/12	64 11/12	65	64 8/12
1966	65 9/12	65	65 2/12	65 2/12
1967	65 10/12	65 1/12	65 4/12	65 4/12
1968	66	65 3/12	65 6/12	65 6/12
1969	66	65 4/12	65 8/12	65 8/12
1970	66 1/12	65 5/12	65 10/12	65 10/12
1971	66 2/12	65 6/12	66	66
1972	66 3/12	65 7/12	66 2/12	66 2/12
1973	66 4/12	65 9/12	66 4/12	66 4/12
1974	66 4/12	65 10/12	66 6/12	66 6/12
1975	66 5/12	65 11/12	66 8/12	66 8/12
1976	66 6/12	66	66 10/12	66 10/12
1977	66 6/12	66 1/12	67	67
1978	66 7/12	66 3/12	67 2/12	67 2/12
1979	66 7/12	66 4/12	67 4/12	67 4/12
1980	66 8/12	66 5/12	67 6/12	67 6/12
1981	66 8/12	66 6/12	67 8/12	67 8/12
1982	66 9/12	66 7/12	67 10/12	67 10/12
1983	66 9/12	66 8/12	68	68
1984	66 10/12	66 9/12	68 2/12	68 2/12
1985	66 10/12	66 10/12	68 4/12	68 4/12
1986	66 11/12	66 11/12	68 6/12	68 6/12
1987	66 11/12	67	68 8/12	68 8/12
1988	66 11/12	67 1/12	68 10/12	68 10/12
1989	67	67 2/12	69	69
1990	67	67 3/12	69 2/12	69 2/12
1991	67 1/12	67 4/12	69 4/12	69 4/12
1992	67 1/12	67 5/12	69 6/12	69 6/12
1993	67 1/12	67 6/12	69 8/12	69 8/12
1994	67 2/12	67 7/12	69 10/12	69 10/12
1995	67 2/12	67 8/12	70	70
1996	67 3/12	67 9/12	70 2/12	70 2/12
1997	67 3/12	67 10/12	70 4/12	70 4/12
1998	67 4/12	67 11/12	70 6/12	70 6/12
1999	67 4/12	68	70 8/12	70 8/12
2000	67 4/12	68 1/12	70 10/12	70 10/12
2001	67 5/12	68 2/12	71	71
2002	67 5/12	68 3/12	71 2/12	71 2/12
2003	67 6/12	68 4/12	71 4/12	71 4/12
2004	67 6/12	68 5/12	71 6/12	71 6/12
2005	67 6/12	68 6/12	71 8/12	71 8/12
2006	67 7/12	68 6/12	71 10/12	71 10/12
2007	67 7/12	68 7/12	72	72
2008	67 7/12	68 8/12	72 2/12	72 2/12
2009	67 8/12	68 9/12	72 4/12	72 4/12
2010	67 8/12	68 10/12	72 6/12	72 6/12
2011	67 9/12	68 11/12	72 8/12	72 8/12
2012	67 9/12	68 11/12	72 10/12	72 10/12
2013	67 9/12	69	73	73
2014	67 10/12	69 1/12	73 2/12	73 2/12
2015	67 10/12	69 2/12	73 4/12	73 4/12
2016	67 11/12	69 3/12	73 6/12	73 6/12
2017	67 11/12	69 3/12	73 8/12	73 8/12
2018	67 11/12	69 4/12	73 10/12	73 10/12
2019	68	69 5/12	74	74
2020	68	69 6/12	74 2/12	74 2/12

1At present time the retirement age of females is lower than for males and depends on the number of their children. It will be subsequently equalised in next decades.
Source: Law No. 155/1995 Coll., author's computations

Conclusions

The paper shows the model calculations of the retirement age needed to ensure that the average period of receipt of a pension is at the level of the last quarter of the expected entire lifetime of people reaching retirement age.

The level of the model value of retirement age understandably depends first and foremost on the prognosis of the mortality development used. For practical purposes it is expected that far more sophisticated methods will be used for predicting the development of mortality, not the simple projections used in this paper. In addition the accuracy of prognoses declines with their length and actual development almost always differs somewhat from the forecast development. The proposal of the Pension Committee therefore envisages that the prognoses would be regularly updated every five years on the basis of the latest available data and that any correction of the retirement age would be carried out only for the generation of people whose age would be between roughly 25-50 years at the moment of making the correction. It would not, then, be possible to change the retirement age of persons for whom, according to current legislation, retirement is only a few years away, nor would the precise value of retirement age be determined in advance for the very young or those not even yet born.

The results of the modelling on the basis of the simple projections of the CZSO or Eurostat, however, confirm in any case that in the generations of persons born up to 1965 there is no reason to lower the existing limit of the retirement age, as in their case the average period of receipt of a pension would actually equal a little more than 25 % of their lives. This applies in particular to women, whose retirement age should rise by 2030 to the level of the retirement age for men or (in the case of women with three or more children, of whom there are relatively few) they should reach this level a few years after 2030.

On the other hand the model calculations indicate that if the average period of receipt of a pension should equal 25 % of a lifetime, then after 2030 (i.e. for the generation of 1966 and younger) there should be a slowing-down of the present tempo of the rise in retirement age. This depends, however, not only on whether the proposal mentioned will be approved, but also on the results of later calculations based on more sophisticated and updated prognoses of the future development of mortality.

Acknowledgment

This article was supported by the Grant Agency of the Czech Republic No. GA ČR 15-13283S under the title *Projection of the Czech Republic Population According to Educational Level and Marital Status*.

References

1. B. Burcin and T. Kučera. Prognóza vývoje obyvatelstva České republiky do roku 2070. In: Bartoňová D. a kol.: Demografická situace České republiky: proměny

a kontexty 1993–2008. Praha: Sociologické nakladatelství SLON, 2010, s. 181–212. ISBN 978-80-7419-024-7.
2. Council Recommendation of 8 July 2014 on the National Reform Programme 2014 of the Czech Republic and delivering a Council opinion on the Convergence Programme of the Czech Republic, 2014. Official Journal of the European Union (2014/C 247/03) http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014H0729(03)&from=CS
3. CZSO (Czech Statistical Office). Projekce obyvatelstva České republiky do roku 2100. https://www.czso.cz/csu/czso/projekce-obyvatelstva-ceske-republiky-do-roku-2100-n-fu4s64b8h4. [Cit. 2015-03-25].
4. Eurostat. 2014. Statistics Database. Database by themes. Population and social conditions. Population projections (proj). EUROPOP2013 - Population projections at national level (proj_13n). Assumptions (proj_13na). Age specific mortality rates by sex (proj_13naasmr). http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=proj_13naasmr&lang=en [Cit. 2014-10-25].
5. Expert Committee on Pension Reform Czech Republic. Final Report on Activities in 2014. http://www.duchodova-komise.cz/wp-content/uploads/2015/02/Final-report-CoE-2014.pdf

6. T. Fiala and J. Langhamrová. Increase of Labor Force of Older Age – Challenge for the Czech Republic in Next Decade. In: 17th International Conference Enterprise and Competitive Environment 2014. [online] Brno, 06.03.2014 – 07.03.2014. Amsterdam: Elsevier, 2014, s. 144–153. Procedia Economics and Finance 12. ISSN 2212-5671. URL: http://www.sciencedirect.com/science/article/pii/S221256711400330X.
7. T. Fiala and M. Miskolczi. Estimation of the number of descendants of pensioners of the given year of births. In: The 8th International Days of Statistics and Economics. [online] Praha, 11.09.2014 – 13.09.2014. Slaný: Melandrium, 2014, s. 405–413. ISBN 978-80-87990-02-5. URL: http://msed.vse.cz/msed_2014/article/453-Fiala-Tomas-paper.pdf.
8. V. Roubíček. Základní problémy obecné a ekonomické demografie. 2. vyd. Praha, 2002: VŠE, 275 s. ISBN 80-245-0288-7.
9. Zákon 155/1995 Sb. O důchodovém pojištění, aktuální znění, Příloha.

A specific semi-markovian dynamic bayesian network estimating residual useful life

Josquin Foulliaron[1], Laurent Bouillaut[1], Patrice Aknin[2] and Anne Barros[3]

[1] University Paris-Est, IFSTTAR, GRETTIA, Champs-sur-Marne F-77447 Marne-la-Vallée Cedex2, France
(E-mail: josquin.foulliaron@ifsttar.fr / laurent.bouillaut@ifsttar.fr)
[2] SNCF - Innovation and Research, 75611 Paris cedex 12, France
(E-mail: patrice.aknin@ifsttar.fr)
[3] NTNU - Department of Production and Quality Engineering, 7491 Trondheim, Norway
(E-mail: anne.barros@ntnu.no)

Abstract. Degradation processes modelling is a key problem to perform any type of reliability study. Indeed, the quality of the computed reliability indicators and prognosis estimations directly depends on this modelling. Mathematical models commonly used in reliability (Markov chains, Gamma processes…) are based on some assumptions that can lead to a loss of information on the degradation dynamic. In many studies, Dynamic Bayesian Networks (DBN) have been proved relevant to represent multicomponent complex systems and to perform reliability studies. In a previous paper, we introduced a degradation model based on DBN named graphical duration model (GDM) in order to represent a wide range of duration models. This paper will introduce a new degradation model based on GDM integrating the concept of conditional sojourn time distributions in order to improve the degradation modelling. It integrates the possibility to take into account several degradation modes together and to adapt the degradation modelling in respect of some new available observations of either the current operation state or the estimated degradation level, to take into account an eventual dynamic change. A comparative study on simulated data between the presented model and the GDM will be performed to show the interest of this new approach.
Keywords: Dynamic Bayesian Networks, Graphical Duration Models, semi-markovian degradation process modelling, Reliability analysis, Residual Useful Life estimation.

1 Introduction

For the last fifty years, the complexity of most of industrial systems has constantly increased. If at best, their failure can lead to a temporary poorer performance of the system, a complete shutdown can also occur, inducing some potentially strong security risks. If the system fails, some components can have to be replaced, making the system unavailable for quite a long time which can be very costly. For these reasons, the research of decision support tools for the reliability analysis has become a key issue.
Many studies already dealt with this topic. Two approaches seems to be mainly used for degradation process modelling: the set of analytic degradation models

16th ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece

derived from the mechanic of the system, Lemaitre and Demorat [1], sometimes quite difficult to validate, and the use of stochastic tools, Aven and Jensen [2]. In this second approach, some models are directly based on probability distributions such as Bertholon model, Bertholon *et al*. [3], Weibull Freitas *et al*. [4] or exponential distributions... These modelling generally aim to focus on the failure time of the system. If one needs to evaluate the temporal behavior of a set of random variables, "dynamic" approaches will be preferred such as stochastic processes (Gamma process, Van Noortwijk [5], Poisson process, Hossain and Dahiya [6]...) or Probabilistic Graphical Models (PGM) such as Neuronal networks, Rajpal *et al*. [7], Petri nets, Volovoi [8], Dynamic Bayesian Networks, Weber and Jouffe [9] ... In this paper, this last formalism was considered. Indeed, since some years, it has been proved as relevant to perform reliability studies since a degradation modelling based on discrete and finite states space is acceptable.

The simplest model of degradation process of a system using the DBN formalism is based on its ability to model a simple Markov chain. Then, this approach implies the strong assumption of geometrically distributed sojourn times in each state. To overcome this limitation, a specific DBN named Graphical duration model (GDM) was proposed, Donat *et al*. [10], to model the degradation of a discrete states system using any kind of discrete sojourn time distribution.

If this modelling provided some interesting results such as Bouillaut *et al*. [11], it assumes that sojourn times elapsed in each state are independent. The existence of several dynamics in the degradation process cannot therefore be identified neither taken into account. In this paper, an extension of the standard semi-markovian GDM modelling is proposed, managing the dependence between the sojourn times through the concept of conditional sojourn time distributions (CSTD). The aim is to be able to build a model that can describe a system whose dynamic is a mixture of some degradations modes and that can be adapted to observe changes in modes.

In the next section, the formalism of DBN and MGD will be briefly introduced and compared. Then, the proposed GDM with conditional sojourn time distributions will be detailed. Finally, a comparative study of the standard markovian approach with DBN and semi-markovian modelling (with GDM and MGD with CSTD) will be proposed for reliability analysis before some conclusions and prospects.

2 Probabilistic Graphical Models, a frame for reliability analysis

2.1 From Bayesian Networks to Dynamic Bayesian Networks

Formally, a Bayesian Network (BN) denoted by $\mathcal{M}$ is defined as a pair $(\mathcal{G}, \{p_n\}_{1 \le n \le N})$ where:

- $\mathcal{G}$=($\boldsymbol{X}$, ε) is a directed acyclic graph giving a qualitative description of the BN. The graph nodes and the associated random variables are both represented by $\boldsymbol{X}$=$\{X_1, \ldots, X_N\}$, with values in $\boldsymbol{\mathcal{X}}$= $\mathcal{X}_1$ x …x $\mathcal{X}_N$. ε is the set of edges encoding the conditional independence relationships among these variables.

- $\{p_n\}_{1\leq n\leq N}$ a set of Conditional Probability Distributions (CPD) associated with the random variables. These distributions aim to quantify the local stochastic behavior of each variable.

Besides, both the qualitative (i.e. $\mathcal{G}$) and quantitative (i.e. $\{p_n\}$) parts of $\mathcal{M}$ can be automatically learnt, if some complete or incomplete data or experts opinions are available, Jensen [12]. Using BN is also particularly interesting because of the easiness for knowledge propagation through the network. Indeed, various inference algorithms allow computing the marginal distribution of any sub-set of variables.

In a dynamic behavior modeling point of view, the time extension of BN provide a convenient formalism to represent discrete sequential systems. Indeed, DBNs are dedicated to model data which are sequentially generated by some complex mechanisms (time-series data, bio-sequences, number of mechanical solicitations before failure…). It is therefore frequently used to model Markov chains. Formally, a DBN is defined by a pair of BN ($\mathcal{M}_{ini}$, $\mathcal{M}^{\rightarrow}$) where:

- $\mathcal{M}_{ini} = \left(\mathcal{G}^{ini}, \left\{p_n^{ini}\right\}_{1\leq n\leq N}\right)$ is a BN modeling the initial distribution of $\boldsymbol{X}$, denoted p^{ini}.

- $\mathcal{M}^{\rightarrow} = \left(\mathcal{G}^{\rightarrow}, \left\{p_{t,n}^{\rightarrow}\right\}_{2\leq t\leq T;\, 1\leq n\leq N}\right)$ defines the transition model of the considered process, i.e. the distribution of $\boldsymbol{X_t}$ knowing $\boldsymbol{X_{t-1}}$, denoted $p^{\rightarrow}$.

Figure 1 introduces a DBN modeling the Markov Chain of the sequence $\boldsymbol{X}$=(X_1, …, X_N) taking its values in the set $\boldsymbol{\mathcal{X}}$. This DBN is described by the pair:

$\left(\mathcal{M}_{ini}, \mathcal{M}^{\rightarrow}\right) = \left(\left(X_1, p_1\right), \left(\mathcal{G}^{\rightarrow}, \mathcal{Q}^{sys}\right)\right)$ where $\mathcal{Q}^{sys}$ denotes the transition matrix of a Markov Chain, quantifying the probability of $X_t|X_{t-1}$.

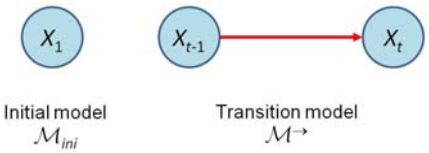

Fig. 1. DBN modelling a Markov Chain

If this approach is perfectly adapted to model the dynamic of systems, it induces a strong assumption on sojourn time distribution in each state of the system. Indeed, as all Markovian approaches, transition rates are assumed constant and, therefore, sojourn times are necessarily geometrically distributed. In many

industrial applications, such an assumption can introduce strong bias in the degradation modeling that cannot be foreseen in a context of reliability based maintenance optimization. To overcome this drawback, a specific DBN, named Graphical Duration model, was proposed and will be briefly introduced in the next paragraphs.

2.2 Graphical Duration Models

The Graphical Duration Model is a specific DBN, using a semi-Markov approach. The main idea is to deal with the couple (X_t, S_t) rather than the single variable $\{X_t\}$ where S_t denotes the remaining time in the current state of X_t. Figure 2 introduces the structure of a DMN modeling a GDM. The solid lines define the basic structure of the GDM; dashed lines indicate optional items and red bold edges characterize dependencies between time slices.

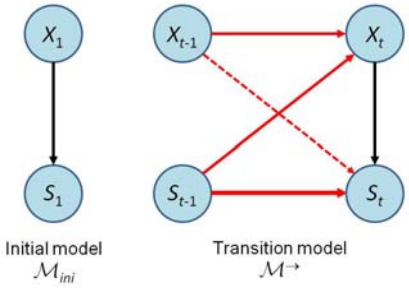

Fig. 2. Specific structure of a DBN modelling a Graphical Duration Model.

A GDM is therefore defined by the pair ($\mathcal{M}_{ini}$, $\mathcal{M}^{\rightarrow}$) with:

- $\mathcal{M}_{ini} = \left(\mathcal{G}^{ini}, (\alpha_1, F_1)\right)$ where α_1 and F_1 denote respectively the initial distribution of X_1 and S_1.
- $\mathcal{M}^{\rightarrow} = \left(\mathcal{G}^{\rightarrow}, (Q^{\rightarrow}, F^{\rightarrow})\right)$ characterized by two transition distributions: $Q^{\rightarrow}$ is the natural states changes distribution and $F^{\rightarrow}$ is the sojourn time distribution, both described by the following equations (1) and (2).

$F^{\rightarrow}$ is the distribution of the remaining sojourn times in the current state X_t. If a natural transition occurs at t (i.e. S_{t-1}=1), this distribution is defined by a given conditional probability distribution F^{sys}. On the other hand, $F^{\rightarrow}$=C where this matrix simply decrement the sojourn time of 1 unit at each iteration, until a natural transition occurs.

$$F^{\rightarrow} = p(S_t = s \mid X_t = x, S_{t-1} = s') = \begin{cases} F^{sys}(x,s) = p(S_t = s \mid, X_t = x, S_{t-1} = 1) \\ C(s',s) = p(S_t = s \mid S_{t-1} = s') \\ \qquad = \begin{cases} 1 \text{ if } s = s'-1 \\ 0 \text{ otherwise} \end{cases} \end{cases} \quad (1)$$

$Q^{\rightarrow}$ is the distribution of state transitions. If a natural transition occurs at t, this distribution is defined by the previously introduced transition matrix Q^{sys}. On the other hand, $Q^{\rightarrow}=I$ enforcing the variable X in the current state until a natural transition occurs.

$$Q^{\rightarrow} = p\left(X_t = x \mid X_{t-1} = x', S_{t-1} = s'\right) = \begin{cases} Q^{sys}(x',x) = p\left(X_t = x \mid, X_{t-1} = x', S_{t-1} = 1\right) \\ I(x',x) = p\left(X_t = x \mid X_{t-1} = x', S_{t-1} > 1\right) \\ \qquad = \begin{cases} 1 \text{ if } x' = x \\ 0 \text{ otherwise} \end{cases} \end{cases} \tag{2}$$

Besides, the structure of a GDM introduced in figure 2 shows the process (X_t) (respectively (S_t)) is not Markovian since $X_{t-1} \not\perp\!\!\!\perp X_{t+1} \mid X_t$ (respectively $S_{t-1} \not\perp\!\!\!\perp S_{t+1} \mid S_t$); where $A \not\perp\!\!\!\perp B$ denotes that variables A and B are not statistically independent. On the other hand, the GDM structure leads to

$$\left(X_{t-1}, S_{t-1}\right) \perp\!\!\!\perp \left(X_{t+1}, S_{t+1}\right) \mid \left(X_t, S_t\right) \tag{3}$$

So, the set (X_t, S_t) engendered by a GDM is Markovian, despite (X_t) is not.

On the practical point of view, this approach allows specifying arbitrary state sojourn time distributions by contrast with a classic Markovian framework in which all durations have to be exponentially distributed. This modeling is therefore particularly interesting as soon as the question is to capture the behavior of a given system subjected to a particular context and a complex degradation distribution. More details on this GDM (quantitative description, optional context description …) can be found in Donat *et al.* [10].

As an illustration of the contribution of GDM for reliability analysis, the simple and "standard" academic 3 serial-parallel components system will be considered. C_2 and C_3 are parallel components, taking their values in {*ok*, *failed*} whereas the serial component C_1 takes its values in {*ok*, *small defect*, *failed*}. The system fails when C_2 and C_3 are simultaneously failed or when C_1 fails.

Figure 3 introduces the sojourn time distributions, considered in this example, for all "not failed" states. Parameters of a DBN modeling a Markov Chain and of GDM are learnt for the three components, using a database with 1000 sojourn times in each state. This learning phase provides transition rates for the standard DBN modeling and discretized sojourn time distributions for the GDM modeling, conditioned by the parameter T_{max}, defined as the higher bound of sojourn times. The settlement of this parameter is a fundamental point in the GDM approach. Indeed, if it is underestimated, the learning of sojourn time distributions can be strongly biased. On the other hand, the complexity of the considered bayesian network can induce algorithmic problems. In this paper, T_{max}=200.

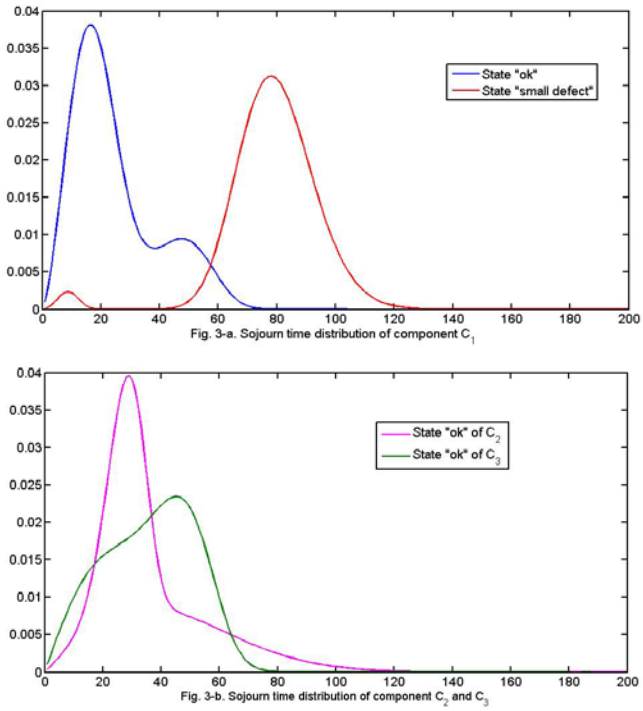

Fig. 3. Sojourn time distributions in non failure states for components C_1, C_2 and C_3.

When all parameters are learnt, all kind of reliability indicators can be easily estimated, such as instantaneous availability, reliability, cumulative distribution function… The following figure introduces the estimation of this last indicator by both modeling "standard" DBN and GDM for the 3 components system.

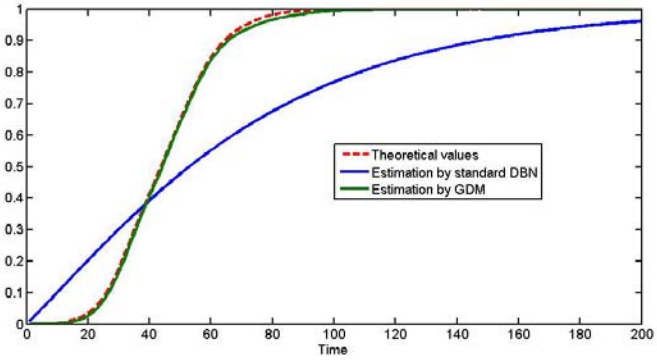

Fig. 4. Sojourn time distributions in non-failure states for components C_1, C_2 and C_3.

The red line introduces the theoretical behavior of the considered reliability indicator. One can note that, due to the stochastic properties of the sojourn time distributions characterizing our system, a Markovian approach can not fit the exact behavior of the system when the GDM provides a good formalism. This first result focuses on the impact of the assumptions made during the degradation modeling on the accuracy of our reliability analysis.
In the next section, an extension of GDM is introduced, allowing taking into account several dynamics in the degradation process modeling. Indeed, as we can note in figure 3, the sojourn times distributions seems to consist in the merging of different dynamics. This is particularly identifiable on figure3-a where two behaviors can be observed. With the standard GDM, this information cannot be taken into account…

3 Introduction of conditional sojourn time distributions in GDM

3.1 Structure and main properties of GDM-CSTD

Figure 5 shows the graphical structure of the DBN modeling a GDM with conditional sojourn time distributions. A variable encoding the current degradation mode, denoted M_t, is added to the couple (X_t,S_t) used by the standard MGD. This change induces updating the transition distributions introduced in (1) and (2).

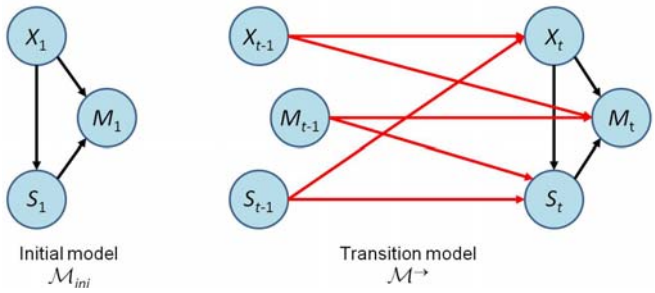

Fig. 5. Specific structure of a DBN modelling a GDM integrating conditional sojourn time distributions.

Since M_{t-1} does not influence X_t, the state transition distribution introduced by (2) will not be modified. The impact of the modes variable on the sojourn time variable S_t leads to the following adaptation of (1).

$$F^{\rightarrow} = p\left(S_t = s \mid X_t = x, S_{t-1} = s', M_{t-1} = m'\right) = \begin{cases} F^{sys}\left(x, s, m'\right) \text{ if } s' = 1 \\ C\left(s', s\right) \text{ if } s' > 1 \end{cases} \quad (4)$$

If no natural transition occurs between t-1 and t, the matrix C decrements the remaining sojourn time by 1 unit whereas if s'=1, the state of the system changes between t-1 and t. Then, the sojourn time in the new current state is

chosen in respect F^{sys}, the probability distribution of the sojourn time for each state, in each mode.

Finally, (5) quantifies the transition of modes. If the state of the system does not change between t-1 and t, the conside red mode at t-1 is conserved using the matrix I. If a state transition occurs, the distribution W^{sys} defines, through the elapsed time in the previous state, the most probable mode for the current state.

$$W^{\rightarrow} = p\left(M_t = m \mid X_{t-1} = x', X_t = x, S_t = s, M_{t-1} = m'\right) = \begin{cases} W^{sys}\left(x', x, s, m', m\right) \text{ if } x' \neq x \\ I\left(x', x, m'\right) \text{ if } x' = x \end{cases} \quad (5)$$

The learning of CSTD consists in three phases. First, the number of modes, denoted n_m, has to be determined. This can be done either by expert's advice or using a criterion such as BIC - *Bayesian Information Criteron*, Schwarze [13], that will determine, through a return of experience (REX) database D_X, the optimal number of mixtures. Then, through D_X, the EM algorithm, Dempster *et al*. [14], is used to estimate the sojourn time distributions in the initial state for each of the n_m modes. This learning also provides a segmentation of D_X in n_m sub-bases dedicated to each mode, denoted D_X^m with $m \in [1..n_m]$. Finally, the sojourn time distributions for each mode in all other states are learnt using the right D_X^m.

To illustrate this modeling, a four states system is considered, taking its values in $\mathcal{X}$={*ok*, *state2*, *state3*, *failed*}, with two degradation modes, periodically observed (T_{obs}). To set a 1000 trajectories learning database, each sojourn time distribution follows a Weibull distribution that parameters are under mentioned:

- *ok*: mode 1~$\mathcal{W}$(2,15) and mode 2~$\mathcal{W}$(6,33)
- *state2*: mode 1~$\mathcal{W}$(6,10) and mode 2~$\mathcal{W}$(9,25)
- *state3*: mode 1~$\mathcal{W}$(6,5) and mode 2~$\mathcal{W}$(15,15)

Then, considering the previously introduced learning procedure, conditional probability distributions are learnt from the sampled database for both MGD and MGD-CSTD approaches. This learning phase underlined the bias that might be introduced in the estimation of the failure time of a periodically observed system when the coexistence of several dynamics in the degradation process is not taken into account. Indeed, if these sojourn time distributions are used to estimate the remaining useful life of our system, in the standard MGD approach, the observation of a short sojourn time in the first state does not impact the estimation of the sojourn time in the next states. The following subsection will introduce some illustrative results on this point.

3.2 Estimation of the remaining useful life of periodically observed systems

In this last illustration, the 4 states system introduced in section 3.1 will be considered with a T_{obs} periodic monitoring, providing the current state of the system each 5 time step.

Aiming to estimate the remaining useful life (RUL) of the system and to update it wh en a n ew ob servation is av ailable, the structure of th e MGD- CSTD introduced in figure 5 was adapted, adding two variables D_t and δ_t that represent respectively the diagnosis of X_t by t he monitoring device, t aking i ts val ues i n $\mathcal{X} \cup \{\varnothing\}$ and an activation variable controlling if the diagnosis is active or not.

In t he c onsidered e xample, δ_t is T_{obs} peri odically activated. In t he other cases, the monitoring device is no t active and D_t returns no information on X_t through the state $\varnothing$. Figure 6 introduces the structure of this new DBN.

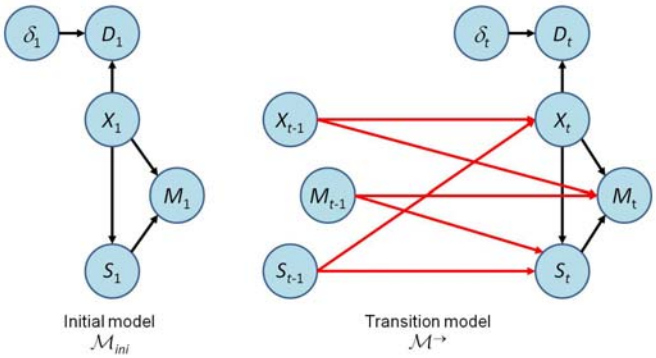

Fig. 6. Structure of the DBN used to estimate the RUL of a periodically observed system with a GDM-CSTD approach.

During th e in itialization p hase, a sojou rn time in state *ok* i s obt ained with F_1, determining the m ost p robable cu rrent m ode m_1. T hen, knowing m_1, s ojourn time di stributions (or co nditional s ojourn t ime di stribution i n t he MGD-CSTD approach) in states *state2* a nd *state3* (con tained in F^{sys}) p rovide t he in itial estimation of the RUL.

When a new diagnosis is av ailable, if a natural tran sition is observed, in th e MGD-CSTD a pproach, the c urrent m ode is event ually corrected i n res pect of the most probable mode knowing th e elap sed time in th e previous state. Th en, with the same reasoning that in the initialization phase, the RUL is uploaded.

Before tryin g to in tegrate this RUL es timation algo rithm in a wid er specific DBN, the proposed methodology has to be briefly evaluated. This was the aim of the results introduced in this end of paper.

Figure 7 i ntroduces two illustrativ e re sults, und erlying t he behavior of MGD and MGD-CSTD a pproaches for t he R UL est imation of t he periodically observed 4 states system.

The first on e focuses on the in terest of using th e cond itional so journ tim e distributions concept. Indeed, one can observe that the information contained in the de gradation m ode (a pproach MGD- CSTD) allows using a m ore accurate sojourn time range to estimate the RUL. This is the reason why the estimation is closer to the real value of the remaining useful life that when the standard MGD is considered.

In the second draw, another interest of MGD-CSTD is underlined. Indeed, in this run, the MGD based estimation looks better in the first observations since it is closer to the real RUL. The explanation of this unusual situation is that, during the initial phase, the mode m_1 was chosen when the right one was m_2. For this reason, the wrong part of the sojourn times was used in the MGD-CSTD approach, explaining the less precise estimations. We can note that at t=35, the situation changes. Between 30 and 35, the state turns from *ok* to *state2*. Then, knowing the elapsed time in *ok*, the algorithm is able to correct its initial mistake by correcting the most probable mode. Then, the MGD-CSTD uses the right part of the sojourn time range while the MGD approach still work with the complete sojourn time domain.

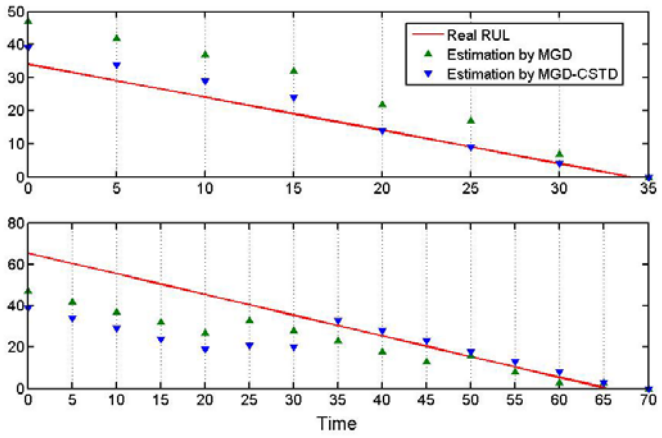

Fig. 7. Examples of RUL estimations using GDM and GDM-CSTD approaches.

To be complete, we have to point on some possible situations that will have to be investigated in further works. Figure 8 introduces what can happen when the sojourn time in a given state is located exactly in the intersection range of two modes. In that example, the mode was correctly initialized in m_1. Then the first estimations with MGD-CSTD are very satisfactory. But, when the transition to state *state2* is detected at t=30, the probability of such a sojourn time in state *ok* for mode m_1 is so weak that an inappropriate mode change is adopted. Then, during the two next observations, the RUL estimation by MGD-CSTD is really poor. Fortunately, in that case, the transition *state2-state3* allows correcting the mode and therefore improving the last RUL estimation.

To give a better idea of the global behavior of the proposed approach, figure 9 introduces the RUL estimation error in respect of time for the trajectories of the considered database obtained with the MGD-CSTD approach, respectively for the 500 trajectories in mode 2 and 93.2% of the 500 trajectories in mode 1. The considered time indices in these drawings correspond to the number of observation before the system fails.

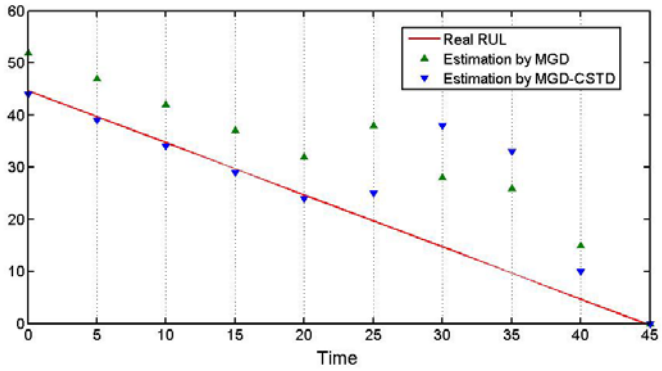

Fig. 8. Examples of RUL estimations using GDM and GDM-CSTD approaches - Unfavourable case.

One can note that, for mode 2 trajectories, the initial estimation of the RUL is quite poor. This is due to the fact that, in the current version of the algorithm, the variable M_t is systematically initialized in mode 1. Then, the first estimations are based on the wrong sojourn time distributions, inducing strong errors. In most of cases, when the first natural state transition is observed, the most probable degradation mode is uploaded and the RUL estimation is improved.

We also can note than the confidence in the estimation increases when the failure time comes and the RUL estimation becomes quite informative. Only very few trajectories are long enough to have more than 14 observations, this is the reason why there are a weak number of points for temporal indices upper than 15.

For mode 1 trajectories, i.e. faster degradation scenarios, we can observe the same global behavior of the RUL estimation algorithm. Nevertheless, this case also illustrates the drawback of the proposed approach, introduced through figure 8 that might be improved in further works. Globally, 34 of the 500 trajectories in mode 1 have at least one sojourn time in the intersection range of two modes, inducing a wrong estimation of the most probable degradation mode, such a way that the adaptation ability of the algorithm, underlined in figure 8, cannot process in these cases and the RUL estimation is absolutely unusable.

Nevertheless, even if the proposed algorithm shows some interesting adaptation abilities when mode estimation errors occur, this final example points out some improvements that will have to be investigated in further works, to make more robust the proposed approach.

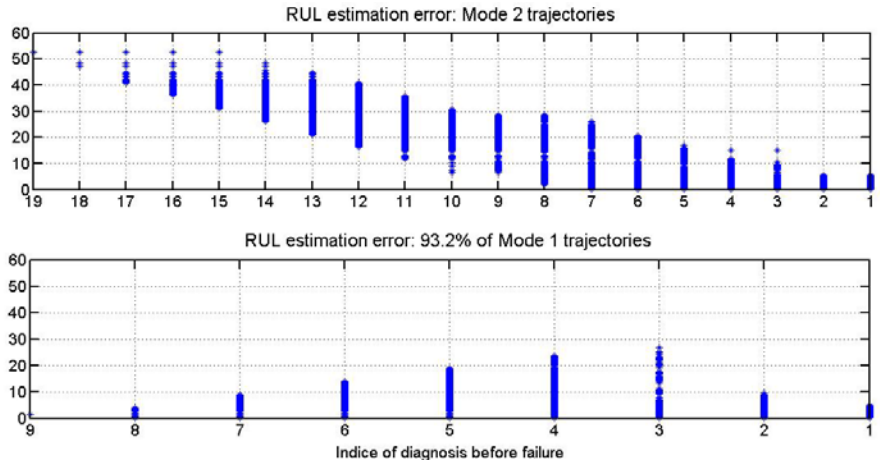

Fig. 9. Global behaviour of the RUL estimation algorithm. Estimation error for each degradation mode in respect of time

Conclusions

In this paper, the formalism of Probabilistic Graphical Models was investigated to determine degradation process modeling for systems with discrete and finite states space. If the "classic" Markovian approach consisting in the modeling of a Markov Chain by a Dynamic Bayesian Network induces necessarily the assumption of sojourn times in each state exponentially distributed, a semi-Markovian approach was proposed using a specific DBN structure, named Graphical Duration Model, that allows considering all kind of sojourn time distribution without any assumption on the stochastic properties of the degradation process.

In this approach, sojourn times in each state are supposed independent. In some applications, especially when the degradation process is the merging of several dynamics, such an assumption can induce strong bias in the estimation of reliability indicators. To illustrate this problem, an example was introduced dealing with the estimation of the remaining useful life of a periodically observed system.

Then, an extension of the standard GDM approach was proposed, integrating the notion of conditional sojourn time distributions by a new random variable managing the most appropriate degradation mode, knowing the elapsed time in the previous state. Then, both MGD and MGD-CSTD were used and compared for the estimation of the remaining useful life of a multi-states system. If the MGD-CSTD approach offers a very interesting global behavior, its main drawback lies in the existence of intersection ranges for several modes, inducing some potential wrong degradation mode estimations and, in the worst cases, an inability of the algorithm to readjust the RUL estimation before the failure time.

Some improvements of the proposed algorithm are currently in progress on this key point.

Acknowledgment

This study is part of the project DIADEM ANR-13-TDMO-04, supported by the French National Agency for Research (ANR), in partnership with University of Technology of Troye, Faiveley Transport, Keolis Rennes and University of technology of Compiègne.

References

1. J. Lemaitre, and R. Demorat, Engineering Damage mechanics, Berlin Springer, 2005
2. T. Aven, J. Jensen, Stochastic models in reliability. Stochastic modelling and applied probability, 1999.
3. H. Bertholon, N. Bousquet, and G. Celeux, An alternative competing risk model to the Weibull distribution for modelling aging in lifetime data analysis. Lifetime Data Analysis, 12(4), pp. 481-504, 2006.
4. M. Freitas, M. de Toledo, E. Colosimo, and M. Pires, Using degradation data to assess reliability: a case study on train wheel degradation. Quality and Reliability Engineering International, 25(5), pp. 607-629, 2009.
5. J. Van Noortwijk, A survey of the application of gamma processes in maintenance. Reliability Engineering & System Safety, 94(1), pp. 2-21, 2009.
6. S. Hossain and R. Dahiya, Estimating the parameters of a non-homogeneous Poisson process model for software reliability. IEEE Transactions on Reliability, 42(4), pp. 604-612, 1993.
7. P.S. Rajpal, K.S. Shishodia and G.S. Sekhon, An artificial neural network for modeling reliability, availability and maintainability of a repairable system, Reliability Engineering and System Safety, 91(7), pp. 809-819, 2006.
8. V. Volovoi, Modeling of System Reliability Using Petri Nets with Aging Tokens, Reliability Engineering and System Safety, 84, pp. 149-161, 2004.
9. P. Weber and L. Jouffe, Reliability modelling with dynamic bayesian networks. In proceedings of the 5th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, 2003.
10. R. Donat, P. Leray, L. Bouillaut, and P. Aknin, A dynamic Bayesian network to represent discrete duration models, Neurocomputing, 73(46), pp. 570-577, 2010.
11. L. Bouillaut, O. François, and S. Dubois, A Bayesian network to evaluate underground rails maintenance strategies in an automation context. Proceedings of the Institution of Mechanical Engineers, Part O, Journal of Risk and Reliability, 227(4), pp.411-424, 2013.
12. F.V. Jensen, An introduction to Bayesian networks. UCL Press, 1996.
13. G. Schwarz, Estimating the dimension of a model, Annals of Statistics, 6(2), pp. 461-464, 1978.
14. A. Dempster, N. Laird and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B(39), pp. 1-38, 1977.

On Seasonal Demand Impact on Availability of Aging Multi-state Manufacturing System

Ilia Frenkel[1], Lev Khvatskin[1] and Anatoly Lisnianski[2]

[1] Center for Reliability and Risk Management, Industrial Engineering and Management Department, SCE-Shamoon College of Engineering, Beer-Sheva, Israel
(E-mail: iliaf@sce.ac.il / khvat@sce.ac.il)

[2] 2The Israel Electric Corporation Ltd. P. O. Box 10, Bait Amir, Haifa 3100, Israel
(E-mail: anatoly-l@iec.co.il)

Abstract. We present availability assessment of multi-state aging industrial animal food additives production manufacturing system and investigate an impact of seasonal demand. In order to determine the system availability we constructed Markov models, representing the various production levels of each element and sub-system in the manufacturing system. Some elements in the system have an aging property. The entire system can be represented as Markov model with 48 different states expressing the different performance levels of the entire process. The production demand is described as two level seasonal Markov model, typical for such production process. The entire Markov model is described as system with 96 differential equations, solution of which is complicated problem. To overcome this obstacle we propose an application of the *Lz*-transform method for availability assessment of aging multi-state system (MSS) and its manufacturing capability.

We demonstrated that the suggested method can be implemented in engineering decision making and construction of various MSS aging systems related to requirements, availability and production.

Keywords: multi-state system; Markov model; *Lz*-transform method; discrete-state continuous-time stochastic process; Availability.

1 Introduction

In this paper we examine a manufacturing production system for animal food additives. The system consists of four elements: reactors and filters. The nominal productivity of the entire system is 600 ton/year.

Due to the system's nature, a fault in a single unit has only partial effect on the entire performance: it only reduces the system's productivity. Therefore, the production system can be assessed as a multi-state system (MSS), where in general both the entire system and its components have an arbitrary predetermined number of states that corresponds to different performance rates (Lisnianski and Levitin [4], Lisnianski et al. [3], Natvig [5]). The performance rate of the system at any instant *t*

16th ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece

is interpreted as a discrete-state continuous-time stochastic process. Such production systems are characterized by numerous states even in relatively simple cases. Therefore, using Markov methods for building the model and finding the solution for the corresponding system of differential equations is rather challenging.

In recent years for dynamic MSS reliability analysis a special technic named L_Z -transform, has been introduced (Lisnianski [2], Lisnianski et al. [3]) for discrete-state continuous-time Markov processes. L_Z –transform extends application of powerful universal generating function technique (Lisnianski [2], Ushalov [6]) to reliability analysis of aging MSS.

In the presented paper, the L_Z -transform is applied to a real production system and its performance is analyzed. It is shown that L_Z -transform application dramatically simplifies the performance computation for such systems in contrast to the straightforward Markov method.

2 Brief Description of the *Lz*-transform Method

In this paper the L_Z-transform method is implemented for the performance determination for MSS manufacturing aging system. The method was introduced by Lisnianski [2] where one can find its detailed description and corresponding mathematical proofs. Briefly, the description of the method is as follows.

We consider a discrete-state continuous-time (DSCT) Markov process $X(t) \in \{x_1,...,x_K\}$, which has K possible states i, (i=1,…,K) where the performance level associated with any state i is x_i. This Markov process is completely defined by the set of possible states $\mathbf{x} = \{x_1, x_2,..., x_K\}$, the transition intensities matrix depending on time $A = (a_{ij}(t)), i, j = 1,2,...,K$ and by the initial states probability distribution given by $\mathbf{p}_0 = [p_{10} = \Pr\{X(0) = x_1\},..., p_{K0} = \Pr\{X(0) = x_K\}]$.

According to [6], the L_Z-transform of a DSCT Markov process $X(t)$ is defined by the following expression

$$L_z\{X(t)\} = \sum_{i=1}^{K} p_i(t) z^{g_i} ,$$

where $p_i(t)$ is the probability that the process is in state i at time instant $t \geq 0$ for a given initial states probability distribution $\mathbf{p}_0$, g_i is the performance level of state i, and z is a complex variable.

In general, any element j in MSS can have k_j different states corresponding to different performance, represented by the set $\mathbf{g}_j = \{g_{j1},..., g_{jk_j}\}$, where g_{ji} is the

performance rate of element j in the state i, $i \in \{1,2,...,k_j\}$, and $j \in \{1,...n\}$, where n is the number of elements in the MSS.

According to L_Z-transform method at first stage, a Markov model of stochastic process should be built for each multi-state element in MSS. Based on this model, state probabilities $p_{ji}(t) = \Pr\{G_j(t) = g_{ji}\}$, $i \in \{1,...,k_j\}$ for every MSS's element can be obtained as a solution of the corresponding system of differential equations under the given initial conditions. These probabilities define output stochastic process $G_j(t)$ for each element j in the MSS. Then, individual L_Z-transform for each element j should be found

$$L_z\{G_j(t)\} = \sum_{i=1}^{k_j} p_{ji}(t) z^{g_{ji}},\ j = 1,...,n$$

At the next stage based on previously determined L_Z-transform for each element j and system structure function f, given by $G(t) = f(G_1(t),...,G_n(t))$, L_Z-transform of the output stochastic process for the entire MSS should be defined. Using Ushakov's operator Ω_f (Ushakov [6]) over all L_Z-transforms of individual elements one can obtain the resulting L_Z-transform $L_z\{G(t)\}$ associated with output performance stochastic process $G(t)$ of the entire MSS:

$$L_z\{G(t)\} = \Omega_f\left\{L_z[G_1(t)],...,L_z[G_n(t)]\right\}$$

The resulting L_Z transform is associated with the output performance stochastic process for the entire MSS:

$$L_z\{G(t)\} = \sum_{k=1}^{K} p_k(t) z^{g_k}$$

and MSS instantaneous availability can be easily derived from the resulting L_Z-transform in the following form:

$$A(t) = \sum_{g_k > 0} p_k(t)$$

In other words, in order to find MSS's mean instantaneous availability one should summarize all probabilities in L_Z-transform from terms where powers of z are greater to zero.

3 Multi-state Model of the Manufacturing Production System in Factory for Animal Food Additives

3.1 System Description

Detailed description of the production system is presented in Frenkel *et al*. [1]. The system consists of connected in series 2 reactor's sub-systems and 2 filter elements. The nominal performance of the whole system is 600 ton/year.

Both reactor subsystems have the nominal performance of 600 ton/year. The first reactor's system is 3 levels system with the following capacities: fully operation state with capacity 600 ton/year, degraded state with capacity 300 ton/year and total failure, corresponding to zero capacity. The second reactor's subsystem is also multi-level system with fully operation state with 600 ton/year capacity and states of partial failures corresponding to capacities 400 and 200 ton/year, and a total failure with capacity 0.

Filters can be in one of two states: a fully operational state with a capacity load of 600 ton per year and a state of total failure corresponding to a capacity of 0. Both filters are elements possess the aging property.

3.2 Reactor's Subsystem No. 1

Figure 1 presents the state-transition diagram of the MSS reactor's subsystem No. 1.

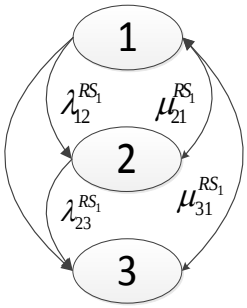

Figure 1. State-transitions diagram of the MSS reactor's subsystem No. 1

Using the state-transitions diagram (Fig. 1) in accordance with the Markov method we build the following system of differential equations for reactor's subsystem No. 1:

$$\begin{cases} \dfrac{dp_1^{RS_1}(t)}{dt} = -\left(\lambda_{12}^{RS_1} + \lambda_{13}^{RS_1}\right) p_1^{RS_1}(t) + \mu_{21}^{RS_1} p_2^{RS_1}(t) + \mu_{31}^{RS_1} p_3^{RS_1}(t); \\ \dfrac{dp_2^{RS_1}(t)}{dt} = \lambda_{12}^{RS_1} p_1^{RS_1}(t) - \left(\lambda_{23}^{RS_1} + \mu_{21}^{RS_1}\right) p_2^{RS_1}(t); \\ \dfrac{dp_3^{RS_1}(t)}{dt} = \lambda_{13}^{RS_1} p_1^{RS_1}(t) + \lambda_{23}^{RS_1} p_2^{RS_1}(t) - \mu_{31}^{RS_1} p_3^{RS_1}(t). \end{cases}$$

Initial conditions are: $p_1^{RS_1}(0) = 1;\ p_2^{RS_1}(0) = p_3^{RS_1}(0) = 0.$

A numerical solution for probabilities $p_i^{RS_1}(t), i = 1,2,3$ can be obtained for this system of differential equations using MATLAB®. Therefore, for reactor's Subsystem No. 1 we can obtain the following output performance stochastic processes:

$$\begin{cases} \mathbf{g}^{RS_1} = \left\{ g_1^{RS_1}, g_2^{RS_1}, g_3^{RS_1} \right\} = \{600, 300, 0\}, \\ \mathbf{p}^{RS_1}(t) = \left\{ p_1^{RS_1}(t), p_2^{RS_1}(t), p_3^{RS_1}(t) \right\}. \end{cases}$$

Having the sets $\mathbf{g}^{RS_1}, \mathbf{p}^{RS_1}(t)$ one can define for Reactor's Subsystem No.1 Lz-transform, associated with the reactor's output performance stochastic process:

$$\begin{aligned} L_z\left\{G^{RS_1}(t)\right\} &= p_1^{RS_1}(t) z^{g_1^{RS_1}} + p_2^{RS_1}(t) z^{g_2^{RS_1}} + p_3^{RS_1}(t) z^{g_3^{RS_1}} \\ &= p_1^{RS_1}(t) z^{600} + p_2^{RS_1}(t) z^{300} + p_3^{RS_1}(t) z^{0}. \end{aligned}$$

3.3 Reactor's Subsystem No. 2

Figure 2 presents the state-transition diagram of the MSS reactor's subsystem No. 2.

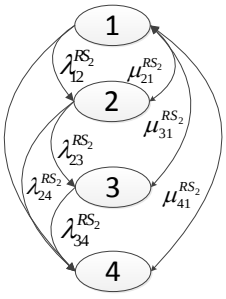

Figure 2. State-transitions diagram of the MSS reactor's subsystem No. 2

Using the state-transitions diagram (Fig. 2) in accordance with the Markov method we build the following system of differential equations for Reactor's Subsystem No. 2:

$$\begin{cases} \dfrac{dp_1^{RS_2}(t)}{dt} = -\left(\lambda_{12}^{RS_2} + \lambda_{14}^{RS_2}\right) p_1^{RS_2}(t) + \mu_{21}^{RS_2} p_2^{RS_2}(t) + \mu_{31}^{RS_2} p_3^{RS_2}(t) + \mu_{41}^{RS_2} p_4^{RS_2}(t); \\ \dfrac{dp_2^{RS_2}(t)}{dt} = \lambda_{12}^{RS_2} p_1^{RS_2}(t) - \left(\lambda_{23}^{RS_2} + \lambda_{24}^{RS_2} + \mu_{21}^{RS_2}\right) p_2^{RS_2}(t); \\ \dfrac{dp_3^{RS_2}(t)}{dt} = \lambda_{23}^{RS_2} p_2^{RS_2}(t) - \left(\lambda_{34}^{RS_2} + \mu_{31}^{RS_2}\right) p_3^{RS_2}(t); \\ \dfrac{dp_4^{RS_2}(t)}{dt} = \lambda_{14}^{RS_2} p_1^{RS_2}(t) + \lambda_{24}^{RS_2} p_2^{RS_2}(t) + \lambda_{34}^{RS_2} p_3^{RS_2}(t) - \mu_{41}^{RS_2} p_4^{RS_2}(t) \end{cases}$$

Initial conditions are: $p_1^{RS_2}(0) = 1;\ p_2^{RS_2}(0) = p_3^{RS_2}(0) = p_4^{RS_2}(0) = 0.$

A numerical solution for probabilities $p_i^{RS_2}(t), i = 1,2,3,4$ can be obtained for this system of differential equations using MATLAB®. Therefore, for Reactor's Subsystem No. 2 we can obtain the following output performance stochastic processes:

$$\begin{cases} \mathbf{g}^{RS_2} = \left\{g_1^{RS_2}, g_2^{RS_2}, g_3^{RS_2}, g_4^{RS_2}\right\} = \{600, 400, 200, 0\}, \\ \mathbf{p}^{RS_2}(t) = \left\{p_1^{RS_2}(t), p_2^{RS_2}(t), p_3^{RS_2}(t), p_4^{RS_2}(t)\right\}. \end{cases}$$

Having the sets $\mathbf{g}^{RS_2}, \mathbf{p}^{RS_2}(t)$ one can define for the Reactor's Subsystem No. 2 L_z-transform, associated with the reactor's output performance stochastic process:

$$\begin{aligned} L_z\left\{G^{RS_2}(t)\right\} &= p_1^{RS_2}(t) z^{g_1^{RS_2}} + p_2^{RS_2}(t) z^{g_2^{RS_2}} + p_3^{RS_2}(t) z^{g_3^{RS_2}} + p_4^{RS_2}(t) z^{g_4^{RS_2}} \\ &= p_1^{RS_2}(t) z^{600} + p_2^{RS_2}(t) z^{400} + p_3^{RS_2}(t) z^{200} + p_4^{RS_2}(t) z^{0}. \end{aligned}$$

3.4 Filter's subsystem

Figure 3 presents the filter's state-transition diagrams.

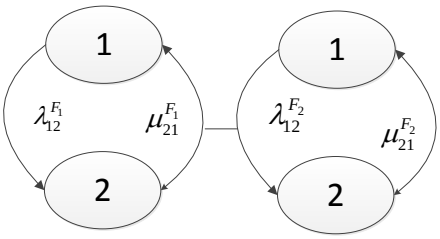

Figure 3. State-transitions diagram of filters No. 1 and 2

Using the state-transitions diagram (Fig. 3) in accordance with the Markov method we build the following system of differential equations for each filter (i=1,2):

$$\begin{cases} \dfrac{dp_1^{F_i}(t)}{dt} = -\lambda^{F_i} p_1^{F_i}(t) + \mu^{F_i} p_2^{F_i}(t), \\ \dfrac{dp_2^{F_i}(t)}{dt} = \lambda^{F_i} p_1^{F_i}(t) - \mu^{F_i} p_2^{F_i}(t). \end{cases} \quad i = 1,2.$$

Initial conditions are: $p_1^{F_i}(0) = 1;\ p_2^{F_i}(0) = 0,\ i = 1,2.$

A numerical solution for probabilities $p_1^{F_i}(t)$ and $p_2^{F_i}(t)$ (i=1,2) can be obtained for each of these 2 systems of differential equations using MATLAB®. Therefore, for each filter we can obtain the following output performance stochastic processes:

$$\begin{cases} \mathbf{g}^{F_i} = \left\{ g_1^{F_i}, g_2^{F_i} \right\} = \{600, 0\}, \\ \mathbf{p}^{F_i}(t) = \left\{ p_1^{F_i}(t), p_2^{F_i}(t) \right\}. \end{cases} \quad i=1,2.$$

Having the sets $\mathbf{g}^{F_i}, \mathbf{p}^{F_i}(t)$ one can define for each filter L_z-transforms, associated with the filter's output performance stochastic process:

$$L_z\left\{G^{F_1}(t)\right\} = p_1^{F_1}(t) z^{g_1^{F_1}} + p_2^{F_1}(t) z^{g_2^{F_1}} = p_1^{F_1}(t) z^{600} + p_2^{F_1}(t) z^0,$$
$$L_z\left\{G^{F_2}(t)\right\} = p_1^{F_2}(t) z^{g_1^{F_2}} + p_2^{F_2}(t) z^{g_2^{F_2}} = p_1^{F_2}(t) z^{600} + p_2^{F_2}(t) z^0$$

3.5 Model for the seasonal demand

The seasonal demand may be described as stochastic demand (Lisnianski[4], Frenkel et al. [2]). Usually, for such systems, demand is seasonally changing: the maximum level is increasing in summer and decreasing in winter.

In this model, the demand is represented as a continuous time Markov process with two states: w_1=500 ton/year is peak level and w_2=250 ton/year is low level. So, corresponding set of demands' levels is as following $\mathbf{w} = \{w_1, w_2\} = \{500, 250\}$. State-transition diagram for Markov process $W(t)$ is presented in Figure 4.

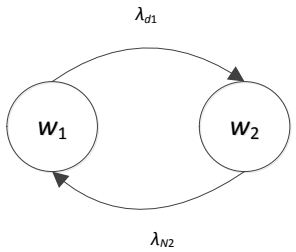

Figure 4. Two level demand model: state-transition diagram

According to the Markov method the systems of differential equations for daily demand is as follows:

$$\begin{cases} \dfrac{dp_{w1}(t)}{dt} = -\lambda_{d1} p_{w1}(t) + \lambda_{N2} p_{w2}(t) \\ \dfrac{dp_{w2}(t)}{dt} = \lambda_{d1} p_{w1}(t) \ - \lambda_{N2} p_{w2}(t) \end{cases}$$

The first state is the system's initial state, so $\mathbf{p}_{w0} = \{p_{w1}(0), p_{w2}(0)\} = \{1, 0\}$.

A numerical solution for probabilities $p_{w1}(t)$ and $p_{w2}(t)$ can be obtained for this system of differential equations using MATLAB®. Therefore, we can obtain the following output performance stochastic process:

$$\begin{cases} \mathbf{W} = \{w_1, w_2\} = \{500, 250\}, \\ \mathbf{p}_w(t) = \{p_{w1}(t), p_{w2}(t)\}. \end{cases}$$

Having the sets $\mathbf{g}_w, \mathbf{p}_w(t)$ one can define for each filter L_z-transforms, associated with the filter's output performance stochastic process:

$$L_z\left\{G_w(t)\right\} = p_{w1}(t)z^{g_{w1}} + p_{w2}(t)z^{g_{w2}} = p_{w1}(t)z^{500} + p_{w2}(t)z^{250}$$

3.6 Multi-state Model for the Production System

All systems' elements are connected in series. So, L_Z-transform, associated with the whole system is:

$$\begin{aligned} L_z\left\{G_s(t)\right\} &= \Omega_{f_{ser}}\left(G^{RS_1}(t), G^{RS_2}(t), G^{F_1}(t), G^{F_2}(t)\right) = \\ &= \Omega_{f_{ser}}\left(p_1^{RS_1}(t)z^{600} + p_2^{RS_1}(t)z^{300} + p_3^{RS_1}(t)z^0\ p_1^{RS_2}(t)z^{600} + p_2^{RS_2}(t)z^{400}\right. \\ &\quad + p_3^{RS_2}(t)z^{200} + p_4^{RS_2}(t)z^0, \\ &\quad \left. p_1^{F_1}(t)z^{600} + p_2^{F_1}(t)z^0, p_1^{F_2}(t)z^{600} + p_2^{F_2}(t)z^0\right). \end{aligned}$$

After simple algebra, where the powers of z are found as minimum values of powers of corresponding terms, the final expression of the whole system's L_Z-transform is of the following form:

$$L_z\left\{G_s(t)\right\} = P_{s1}(t)z^{600} + P_{s2}(t)z^{400} + P_{s3}(t)z^{300} + P_{s4}(t)z^{200} + P_{s5}(t)z^0$$

where

$g_{s1} = 600$ ton/year $\quad P_{s1}(t) = p_1^{RS_1}(t)\,p_1^{RS_2}(t)\,p_1^{F_1}(t)\,p_1^{F_2}(t)$

$g_{s2} = 400$ ton/year $\quad P_{s2}(t) = p_1^{RS_1}(t)\,p_2^{RS_2}(t)\,p_1^{F_1}(t)\,p_1^{F_2}(t)$

$g_{s3} = 300$ ton/year $\quad P_{s3}(t) = p_2^{RS_1}(t)\left(p_1^{RS_2}(t) + p_2^{RS_2}(t)\right)p_1^{F_1}(t)\,p_1^{F_2}(t)$

$g_{s4} = 200$ ton/year $\quad P_{s3}(t) = p_1^{RS_1}(t)\,p_3^{RS_2}(t)\,p_1^{F_1}(t)\,p_1^{F_2}(t)$

$g_{s5} = 0$ ton/year $\quad P_{s5}(t) = p_1^{RS_1}(t)\,p_2^{RS_2}(t) + p_2^{RS_1}(t)\left(p_1^{RS_2}(t) + p_2^{RS_2}(t)\right)p_1^{F_1}(t)\,p_1^{F_2}(t)$

3.7 Availability Computation for Entire Multi-state system

Block diagram for availability computation of the manufacturing systems working under seasonal stochastic demand is presented in the Figure 5.

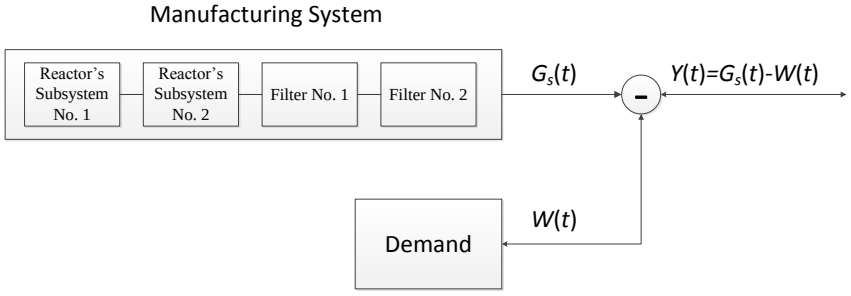

Figure. 5. Block diagram for the MSS availability computation

When the resulting stochastic process $Y(t) = G_s(t) - W(t)$ falls down to level zero such event is treated as a failure. Processes $G_s(t)$, $W(t)$ are independent.

Let's find instantaneous availability for this aging multi-state manufacturing system under the seasonal stochastic demand. .

In according to the L_Z –transform approach we obtain:

$$L_z\left\{G_s(t)\right\} = P_{s1}(t)z^{600} + P_{s2}(t)z^{400} + P_{s3}(t)z^{300} + P_{s4}(t)z^{200} + P_{s5}(t)z^0$$

$$L_z\left\{W(t)\right\} = p_{w1}(t)z^{g_{w1}} + p_{w2}(t)z^{g_{w2}} = p_{w1}(t)z^{500} + p_{w2}(t)z^{250}$$

Now, we have

$$L_z\{Y(t)\} = L_Z\{G_s(t) - W(t)\} = \Omega_{f\text{minus}}\{L_z\{G_s(t)\}, L_z\{W(t)\}$$

$$= \Omega_{f\text{minus}}\left\{P_{s1}(t)z^{600} + P_{s2}(t)z^{400} + P_{s3}(t)z^{300} + P_{s4}(t)z^{200} + P_{s5}(t)z^0,\right.$$

$$\left. p_{w1}(t)z^{500} + p_{w2}(t)z^{250}\right\}$$

$$= P_{s1}(t)p_{w2}(t)z^{350} + P_{s2}(t)p_{w2}(t)z^{150} + P_{s1}(t)p_{w1}(t)z^{100} + P_{s3}(t)p_{w2}(t)z^{50}$$

$$+ P_{s4}(t)p_{w2}(t)z^{-50} + P_{s2}(t)p_{w1}(t)z^{-100} + P_{s5}(t)p_{w2}(t)z^{-250} + P_{s5}(t)p_{w1}(t)z^{500}$$

Based on the last expression we obtain:

$$L_z\{Y(t)\} = P_{Y1}(t)z^{350} + P_{Y2}(t)z^{150} + p_{Y3}(t)z^{100} + P_{Y4}(t)z^{50} + P_{Y5}(t)z^{-50} + P_{Y6}(t)z^{-100} + P_{Y7}(t)z^{-250} + P_{Y8}(t)z^{-500}$$

where

$g_{Y1} = 350$ ton/year $\quad p_{Y1}(t) = P_{s1}(t)p_{w2}(t)$

$g_{Y2} = 150$ ton/year $\quad p_{Y2}(t) = P_{s2}(t)p_{w2}(t)$

$g_{Y3} = 100$ ton/year $\quad p_{Y3}(t) = P_{s1}(t)p_{w1}(t)$

$g_{Y4} = 50$ ton/year, $\quad p_{Y4}(t) = P_{s3}(t)p_{w2}(t)$

$g_{Y5} = -50$ ton/year, $\quad p_{Y5}(t) = P_{s4}(t)p_{w2}(t)$

$g_{Y6} = -100$ ton/year, $\quad p_{Y6}(t) = P_{s2}(t)p_{w1}(t)$

$g_{Y7} = -250$ ton/year, $\quad p_{Y7}(t) = P_{s5}(t)p_{w2}(t)$

$g_{Y8} = -500$ ton/year, $\quad p_{Y8}(t) = P_{s5}(t)p_{w1}(t)$

These two sets

$$\mathbf{g} = \{g_{Y1}, g_{Y2}, g_{Y3}, g_{Y4}, g_{Y5}, g_{Y6}, g_{Y7}, g_{Y8}\}$$

$$\mathbf{p}(t) = \{p_{Y1}(t), p_{Y2}(t), p_{Y3}(t), p_{Y4}(t), p_{Y5}(t), p_{Y6}(t), p_{Y7}(t), p_{Y8}(t)\}$$

define capacities and states probabilities of output performance stochastic process for the entire MSS.

Based on the resulting L_z-transform $L_z\{Y(t)\}$ of the entire MSS, one can obtain the MSS instantaneous availability of the air-conditioning system under seasonal stochastic demand as

$$A(t) = \sum_{g_Y \geq 0} p_i(t) = p_{Y1}(t) + p_{Y2}(t) + p_{Y3}(t) + p_{Y4}(t)$$

4 Availability Calculation

Calculations were performed using the following failure and repair rates.

The failure rates of the first Rector's Subsystem are $\lambda_{12}^{RS_1} = 2$ year^{-1}, $\lambda_{13}^{RS_1} = 2.3$ year^{-1}, $\lambda_{23}^{RS_1} = 4.3$ year^{-1}. The repair rates are $\mu_{21}^{RS_1} = 300$ year^{-1} and $\mu_{31}^{RS_1} = 300$ year^{-1}

The failure rates of the second Rector's Subsystem are $\lambda_{12}^{RS_2} = 2$ year^{-1}, $\lambda_{23}^{RS_2} = 2$ year^{-1}, $\lambda_{14}^{RS_2} = \lambda_{24}^{RS_2} = 2.5$ year^{-1} and $\lambda_{34}^{RS_2} = 4.5$ year^{-1} . The repair rates are $\mu_{21}^{RS_2} = 730$ year^{-1} $\mu_{31}^{RS_2} = 365$ year^{-1} and $\mu_{41}^{RS_2} = 487$ year^{-1} .

The failure rate of each filter is $\lambda^F = 5 + 2t$ year^{-1}. The repair rate of the pump is $\mu^F = 730$ year^{-1}. As one can see the failure rates of both filters are increasing functions of time, these elements possessing the aging property.

Availability calculation of MSS production system is presented in Figure 6. The curves on this figure show that the availability of an aging system decreases with time and down during a year. 3 curves are presented on this figure: availability for seasonal demand and availabilities tor low level and high level constant demands.

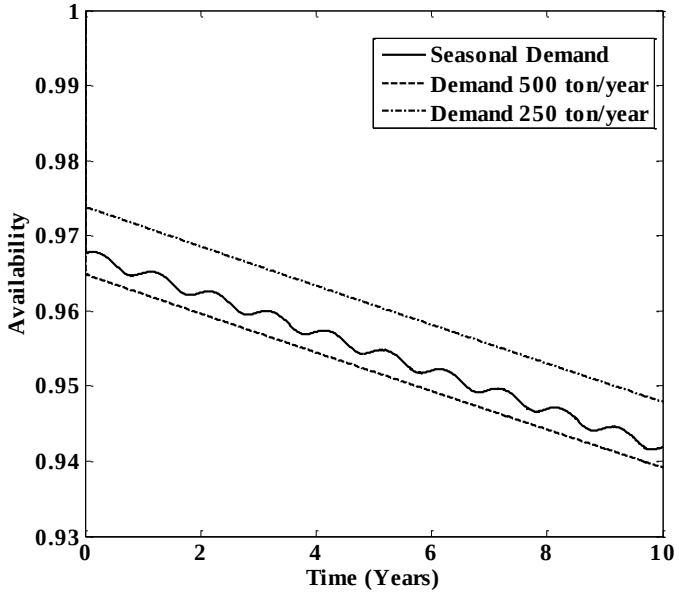

Fig. 6. Compatrison availability of the MSS manufacturing system for different demnd levels

Conclusions

In this paper we applied the L_Z-transform to a practical problem of availability calculation for aging MSS working under stochastic demand. To illustrate, we used as an example a manufacturing system.

We demonstrated that the L_Z-transform method is well formalized and suitable for practical application in reliability engineering for a real-world MSSs analysis. It reinforces engineering decision making and determines a system structure to provide required performance level for complex multi-state aging systems. This method allows dramatic reduction in computation burden in comparison to straightforward Markov method that otherwise would have required building up and solution of a model with enormous number of states.

Acknowledgements

This work was partially supported by the Internal Funding Program of the SCE - Shamoon College of Engineering.

References

1. I. Frenkel, L. Khvatskin, A. Lisnianski, S. Daichman, N. Avraham and O. Zihry. Performance Determination for MSS Manufacturing System by Lz-transform and Stochastic Processes Approach, in Proceedings of the 9th International Conference on Availability, Reliability and Security (ARES2014), Fribourg, Switzerland, September 8-12, 2014, pp 387 - 392, IEEE CPS, 978-1-4799-4223-7/14, DOI 10.1109/ARES.2014.58
2. A. Lisnianski. L_z-transform for a Discrete-state Continuous-time Markov Process and its Application to Multi-state System Reliability, in: Recent Advances in System Reliability. Signatures, Multi-state Systems and Statistical Inference. A. Lisnianski and I. Frenkel, Eds. London: Springer, 2012, 79-95.
3. A. Lisnianski, I., Frenkel and Y. Ding. Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers. Springer, London, 2010.
4. A. Lisnianski and G. Levitin. Multi-state system reliability: assessment, optimization and applications. Singapore: World Scientific, 2003.
5. B. Natvig. Multistate Systems Reliability. Theory with Applications. New York: Wiley, 2011.
6. I. Ushakov. A Universal Generating Function, Soviet Journal of Computer and System Sciences, 24, 37–49. 1986.

Optimal Partition of Markov Models and Automatic Classification of Languages

Jesús E. García[1] and V.A. González-López[2]

[1] Department of Statistics, University of Campinas, R. Sérgio Buarque de Holanda, 651, Campinas (CEP 13083-859), Brazil.
(E-mail: jg@ime.unicamp.br)
[2] Department of Statistics, University of Campinas, R. Sérgio Buarque de Holanda, 651, Campinas (CEP 13083-859), Brazil.
(E-mail: veronica@ime.unicamp.br)

Abstract. In this paper we introduce a new methodology for the problem of automatic classification of languages according to rhythmic features, using speech samples. The problem is to divide the set of languages in subsets with similar rhythmic properties in an automatic way. In other words we look for a partition of the set of languages such that two languages are in the same part of the partition, if and only if, they share the same rhythmic properties. The available dataset consist of 1648 recorded sentences coming from 8 languages. We extract from the speech samples the local energy level on the acoustic signal for two specific frequency bands. Those two energy bands carry information about rhythmic features of the language, according to the results of García *et al.* [7], García *et al.* [12] and García *et al.* [11]. In this way, for each speech sample, we obtain a sequence of energy values. The strategy is to compare the Bayesian information criterion (BIC) computed on the assumption of markovianity, for all possible partitions of the set of languages $\{1, 2, ...8\}$. Assuming that if two samples come from the same source, also follow the same law, which is equivalent to say that if the rhythmic properties of two languages are significantly different, then those languages will be allocated in different parts of the partition of $\{1, 2, ...8\}$. Taking the previous idea in consideration, it was developed an algorithm for the partition selection, under the scope of Partition Markov models (see García and González-López [8]). The resulting partitioning is in agreement with previous results about this problem.
Keywords: Markov models, Bayesian information criterion, Partition of models, Rhythmic classification of languages.

1 Introduction

This paper investigates aspects related to the the rhythmic patterns in speech samples from several languages. In García *et al.* [7] an algorithm is proposed to automatically segment English speech on intervals of vowels and intervals of consonants in function of the energy on some specific bands of frequencies. Ramus *et al.* [14] proposed rhythmic measures based on that segmentation and used them to classify the languages on tree rhythmic classes which correspond

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

to a linguistic conjecture about the rhythm of languages (see Abercrombie [1]). This suggests that it may be possible to use the information about the energy on the frequency bands used by García *et al.* [7] to obtain rhythmic classes. In García *et al.* [11] the same energy bands are used to discriminate between languages and a robust procedure, see García *et al.* [13], it is applied to automatically choose for each language a subset of samples which are similar. We use data from the following eight languages: Catalan, Dutch, English, French, Italian, Japanese, Polish and Spanish. The linguistic conjecture that motivates this study, claims the existence of three rhythmic classes, (a) the stress-timed languages, (b) the syllable-timed languages and (c) the mora-timed languages. Those classes are based in the idea that within each class, different elements cause the temporal organization. The rhythmic type should be correlated with the speech segmentation unit. Speakers of stress-timed languages should have speech segmented in "feet", speakers of syllable-timed languages in "syllables", and speakers of mora-timed languages in "morae". About (a) and (b), Dauer [3] and Ramus *et al.* [14] emphasize two phonologic/phonetic properties. Such characteristics are (i) syllable structure: stress-timed languages have a greater variety of syllable types than syllable-timed languages and (ii) vowel reduction: in stress-timed languages, unstressed syllables usually have a reduced vocalic system. According to (i) and (ii), Spanish, French and Italian should be classified as syllable-timed languages. Dutch and English should be classified as stress-timed languages. The existence of intermediate languages mixing (i) and (ii) was reported in Ramus *et al.* [14]. In fact, Catalan and Polish should be examples. Some studies also show that even in branches of the same language, differences can be found, see Galves *et al.* [6]. And longitudinal studies show that a language can alter their phonological features, see Frota *et al.* [4].
Ramus *et al.* [14] shows evidence that simple statistics based on hand labeled segmentation, in vowels and consonants, of the speech signal could be used to detect rhythmic classes. The limitation of the used approach is that it depends on the segmentation that has to be made by hand by a phoneticist. As a consequence, the sample analyzed was small and also the results are dependent on the phoneticist interpretation. Galves *et al.* [5] proposed a sonority function that assigns to each sentence the mean value (over all the sentence) of a local index of regularity on the spectrogram. The local index of regularity is based on the relative entropy of successive columns of the spectrogram. The sonority function is calculated by an automatized algorithm without any hand labeling, thus avoiding some drawbacks present in Ramus *et al.* [14]. However, the paper does not assess any statistical evidence of difference between the sonority of languages in different rhythmic classes. Cuesta-Albertos *et al.* [2] applied a new Kolmogorov Smirnov test for functional data on the sonority function. They use this test to find statistically significant differences between some of the languages on the different rhythmic classes. This approach discriminates the languages without proposing a model. The main restriction with the sonority approach is that it depends on the relative entropy on the spectrogram of the signal which is very sensitive to small changes in the level of noise on the signal. In the process of collecting acoustic signal, there are several sources which impact in the signal-to-noise ratio, significantly changing the value of the

sonority function. Some changes are generated by external factors at rhythmic properties: different recording instruments, different distance between the microphone and the speaker, natural voice volume of the speakers, etc. For instance, for this reason Cuesta-Albertos *et al.* [2] used just 20 sentences, the same that Ramus used, from a total of around 200 sentences for each language. In contrast, the methodology introduced in the present paper produces, in an automatic way, clusters which are based on bivariate Markov models. Such models fit the time dependence and also fit the dependence between the two energy on the two frequency bands. The mean energy is much less sensitive to noise levels in the recordings than the relative entropy, giving as a result a more robust classification. The clusters obtained confirm the theoretical conjecture and had the advantage compared to Ramus *et al.* [14] and Cuesta-Albertos *et al.* [2] that it does not require pre-selection of sentences. Our fitted models offer a mathematical justification of the linguistic classes established by the conjecture and make no prior assumptions about the rhythm's classes that should be found.

2 The data

The data set consists of 1648 recorded sentences belonging to eight languages, Catalan with 216 sentences, Dutch with 228 sentences, English with 132 sentences, French with 216 sentences, Italian with 216 sentences, Japanese with 212 sentences, Polish with 216 sentences and Spanish with 212 sentences. The sentences have lengths going from 2 to 3.5 seconds, digitalized at 16.000 samples a second (i.e. sample rate of 16 kHz). This data comes from a corpus belonging to the *Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS)*. The corpus includes the 160 sentences analyzed by Ramus *et al.* [14] and Cuesta-Albertos *et al.* [2].
Denote by $\vartheta_t(f)$ the power spectral density at time t and frequency f, which is the square of the coefficient for frequency f of the local Fourier decomposition of the speech signal. The time is discretized in steps of 25 milliseconds and the frequency is discretized in steps of 20 hz. The values of the power spectral density are estimated using a 25 milliseconds Gaussian window.
Fixed a language l we consider the sentence j of length $T_{l,j}$. Given a frequency f we denote by $\vartheta_t^{l,j}(f)$ the power spectral density at time t for that sentence j and language l where $t = 1, \cdots, T_{l,j}$. For each time t we consider the stochastic processes $\chi_1^{l,j}(t) = \sum_{f=80,100,\ldots,800} \vartheta_t^{l,j}(f)$ and $\chi_2^{l,j}(t) = \sum_{f=1500,1520,\ldots,5000} \vartheta_t^{l,j}(f)$ named energies. The definition of the energy bands including the frequencies for the bands were chosen based in previous works about automatic segmentation of speech signal in vowels and consonants, see for example García *et al.* [7].

2.1 Coding the data

For each sentence j from language l and energy band k ($k = 1$ represents the inferior band of energy $\chi_1^{l,j}(t)$ and $k = 2$ represents the superior band of energy

$\chi_2^{l,j}(t)$). We define $Y_t^{l,j,k} = 1$ if $\chi_k^{l,j}(t+1) \geq \chi_k^{l,j}(t)$, and $Y_t^{l,j,k} = 0$ otherwise. Define $Z_{j,t}^l = 2Y_t^{l,j,2} + Y_t^{l,j,1}$.

Remark 1 *The value $Z_{j,t}^l = 0$ means both energies decrease at time $t+1$; $Z_{j,t}^l = 1 (Z_{j,t}^l = 2)$ means that the energy in the inferior(superior) band increases and the energy in the superior(inferior) band decreases at time $t+1$; $Z_{j,t}^l = 3$ means both energies increase at time $t+1$.*

3 Partition Markov models

The Partition Markov Models applied in this paper, were introduced in García and González-López [8]. Those models are generalizations of Variable Length Markov Chains, used to discover the differences in rhythmic features between branches of the Portuguese in Galves *et al.* [6].

Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. Let us call $\mathcal{S} = A^M$ the state space. Denote the string $a_m a_{m+1} \dots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in \mathcal{S}$, $P(a|s) = \text{Prob}(X_t = a | X_{t-M}^{t-1} = s)$. Let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$, for $a \in A$, $L \in \mathcal{L}$, $P(L,a) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s, X_t = a)$, $P(L) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s)$ and $P(a|L) = \frac{P(L,a)}{P(L)}$, with $P(L) > 0$.

Definition 1. Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. We will say that $s, r \in \mathcal{S}$ are equivalent (denoted by $s \sim_p r$) if $P(a|s) = P(a|r)\ \forall a \in A$. For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.

The previous definition allows to define a Markov chain with a "minimal partition", that is the one which respects the equivalence relationship.

Definition 2. let (X_t) be a discrete time, order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$. We will say that (X_t) is a Markov chain with partition $\mathcal{L}$, if this partition is the one defined by the equivalence relationship $\sim_p$ introduced by definition 1.

The set of parameters for a Markov chain over the alphabet A with partition $\mathcal{L}$ can be denoted by $\{P(a|L) : a \in A, L \in \mathcal{L}\}$. Having established the equivalence relationship for a given Markov chain, then we need $(|A| - 1)$ transition probabilities for each part to specify the model. The total number of parameters for the model is $|\mathcal{L}|(|A| - 1)$.

Given a sample x_1^n, $L \in \mathcal{L}, a \in A$, $N(L,a) = \sum_{s \in L} N(s,a)$ and $N(L) = \sum_{s \in L} N(s)$, where the number of occurrences of s in the sample x_1^n is denoted by $N(s)$ and the number of occurrences of s followed by a in the sample x_1^n is denoted by $N(s,a)$.
The model, in this context given by the "optimal partition $\mathcal{L}$", can be selected

consistently, using the Bayesian Information Criterion. This is, the best partition is the one that maximizes

$$\mathrm{BIC}(x_1^n, \mathcal{L}) = \sum_{a \in A, L \in \mathcal{L}} N(L,a) \ln\left(\frac{N(L,a)}{N(L)}\right) - \frac{(|A|-1)|\mathcal{L}|}{2} \ln(n),$$

over the set of partitions of $\mathcal{S}$.

A practical way to choose a model in the family in a consistent manner can be found in García and González-López [8]. Here we propose to use a distance similar to the one defined in the next paragraph in $S_{/\sim_n}$, where $s \sim_n r \iff \frac{N(s,a)}{N(s)} = \frac{N(r,a)}{N(r)} \ \forall a \in A$, where n is the size of the dataset.

Definition 3. Let n be the size of the dataset, for any $s, r \in \mathcal{S}$, $N(\{s,r\},a) = N(s,a) + N(r,a), a \in A$,

$$\begin{aligned} d_n(s,r) = \frac{2}{(|A|-1)\ln(n)} \sum_{a \in A} \Bigg\{ & N(s,a) \ln\left(\frac{N(s,a)}{N(s)}\right) \\ & + N(r,a) \ln\left(\frac{N(r,a)}{N(r)}\right) \\ & - \left(N(\{s,r\},a) \ln\left(\frac{N(\{s,r\},a)}{N(s)+N(r)}\right)\right) \Bigg\}, \end{aligned}$$

d_n can be generalized to subsets of $\mathcal{S}$ and it has the property of being equivalent to the BIC criterion, to decide if $s \sim_p r$ for any $s, r \in \mathcal{S}$ (see García and González-López [8]).

Remark 2 *i. As a consequence of Theorem 2.1 proved in García and González-López [8], if (X_t) is a discrete time, order M Markov chain on a finite alphabet A and x_1^n is a sample of the process, then for n large enough, for each $s, r \in \mathcal{S}$, $d_n(r,s) < 1$ iff s and r belong to the same class.*

ii. The algorithm introduced in García and González-López [9] returns the true partition for the source, this means that under the assumptions of Theorem 2.1 García and González-López [8], $\hat{\mathcal{L}}_n$ given by the algorithm converges almost surely eventually to $\mathcal{L}$, where $\mathcal{L}$ is the partition of $\mathcal{S}$ defined by the equivalence relationship introduced in definition 2.

4 Partition of the set of Markov models corresponding to the languages

Consider the stochastic processes $Z_1, Z_2, ..., Z_8$ corresponding to the eight languages: Catalan, Dutch, English, French, Italian, Japanese, Polish and Spanish. With sample $(z_{i,l})_{l=1}^{n_i}$ of size $n_i, i = 1, \ldots, 8$. Following the codification given in section 2.1, each sample will be composed by the concatenation of symbols from $A = \{0,1,2,3\}$. The value of the order M considered here was 4, based on previous works that investigate similar data, see for example García

et al. [11].
For $\{i_1, ..., i_k\} \subseteq \{1, 2, ..., 8\}$, $s \in \mathcal{S}$ and $a \in A$. Define the counting quantities,

$$N_{\{i_1,...,i_k\}}(s) = \sum_{j=1}^{k} N_{i_j}(s) \text{ and } N_{\{i_1,...,i_k\}}(s,a) = \sum_{j=1}^{k} N_{i_j}(s,a).$$

Where $N_{i_j}(s)$ is the number of occurrences of s in the sample of the i_j-th language, $(z_{i_j,l})_{l=1}^{n_{i_j}}$ and $N_{i_j}(s,a)$ is the number of occurrences of s followed by a. Assuming that the data collection is made up of independent speech samples (which is the case treated here, as each set of samples corresponds to different languages), the BIC under the assumption of independence and identical distribution for $Z_{i_1}, Z_{i_2}, ..., Z_{i_k}$ given an arbitrary partition $\mathcal{L}$ is
$\text{BIC}\Big((z_{i_1,l})_{l=1}^{n_{i_1}}, ..., (z_{i_k,l})_{l=1}^{n_{i_k}}, \mathcal{L}\Big) =$

$$\sum_{a\in A, L\in\mathcal{L}} N_{\{i_1,...,i_k\}}(L,a) \ln\Big(\frac{N_{\{i_1,...,i_k\}}(L,a)}{N_{\{i_1,...,i_k\}}(L)}\Big) - \frac{(|A|-1)}{2}|\mathcal{L}| \ln(n_{i_1} + ... + n_{i_k}). \quad (1)$$

Assumption 1 *$\mathcal{M} = \{M_1, M_2, ..., M_m\}$ is a partition of $\{1, 2, ..., 8\}$ such that, for each $l \in \{1, 2, ..., m\}$ and for each $i, j \in \{1, 2, ..., 8\}$*

$$i, j \in M_l \iff Z_i =^d Z_j \textit{(they have the same distribution law).}$$

Then, we can conclude that the BIC criterion computed from a specific $\mathcal{M}$ of m elements will be expressed in terms of the m parts of the state space, following definition 2, say

$$\text{BIC}\Big((z_{1,l})_{l=1}^{n_1}, ..., (z_{8,l})_{l=1}^{n_8}, \mathcal{M}, \{\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_m\}\Big), \quad (2)$$

to remark its dependence from $\mathcal{M}$. Where $\mathcal{L}_i$ is the partition of the state space associated to $M_i, i = 1, \ldots, m$.
Under the assumption 1, (2) will be

$$\text{BIC}\Big((z_{1,l})_{l=1}^{n_1}, ..., (z_{8,l})_{l=1}^{n_8}, \mathcal{M}, \{\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_m\}\Big) = \sum_{i=1}^{m} \text{BIC}\Big(\{(z_{j,l})_{l=1}^{n_j}\}_{j\in M_i}, \mathcal{L}_i\Big).$$

Where each term of the sum on the right is defined by equation (1).

Maximizing $\text{BIC}\Big((z_{1,l})_{l=1}^{n_1}, ..., (z_{8,l})_{l=1}^{n_8}, \mathcal{M}, \{\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_m\}\Big)$ we get the optimal partition of the set of languages $\{1, 2, ..., 8\}$.

4.1 The strategy

For a fixed $\mathcal{M} = \{M_1, M_2, ..., M_m\}$, define

$$\text{BIC}\Big((z_{1,l})_{l=1}^{n_1}, ..., (z_{8,l})_{l=1}^{n_8}, \mathcal{M}\Big) = \max_{\{\mathcal{L}_1,..,\mathcal{L}_m\}\in\mathcal{P}} \sum_{i=1}^{m} \text{BIC}\Big(\{(z_{j,l})_{l=1}^{n_j}\}_{j\in M_i}, \mathcal{L}_i\Big).$$

Where $\mathcal{P}$ is the set of partitions of $\mathcal{S}$.
To find this maximum it is necessary to maximize each term in the sum. Then, for each $i = 1, 2, ..., m$ we find the partition $\mathcal{L}_i$ of $\mathcal{S}$ which maximizes (in $\mathcal{L}$)

$$\text{BIC}\left(\{(z_{j,l})_{l=1}^{n_j}\}_{j\in M_i}, \mathcal{L}\right).$$

This can be done on an efficient way using a distance similar to the introduced on definition 3, and defined as follows.

Definition 4. Let M_i be one part of $\mathcal{M}$, for any $s, r \in \mathcal{S}$,

$$\begin{aligned} d_{M_i}(s,r) = \frac{2}{(|A|-1)\ln(n)} \sum_{a\in A} \Bigg\{ & N_{M_i}(s,a) \ln\left(\frac{N_{M_i}(s,a)}{N_{M_i}(s)}\right) \\ & + N_{M_i}(r,a) \ln\left(\frac{N_{M_i}(r,a)}{N_{M_i}(r)}\right) \\ & - \left(N_{M_i}(\{s,r\},a) \ln\left(\frac{N_{M_i}(\{s,r\},a)}{N_{M_i}(s)+N_{M_i}(r)}\right)\right)\Bigg\}, \end{aligned}$$

where $n = \sum_{j\in M_i} n_j$, $N_{M_i}(\{s,r\},a) = N_{M_i}(s,a) + N_{M_i}(r,a)$, $N_{M_i}(s,a) = \sum_{j\in M_i} N_j(s,a)$ and $N_{M_i}(s) = \sum_{j\in M_i} N_j(s)$, with $a \in A$.

To obtain $\mathcal{L}_i$ we can use a clustering algorithm using the distance d_{M_i}, in this work, the following algorithm is proposed.

Algorithm 1 *(Markov partition model selection algorithm for the set of samples in M_i)*
Input: $d_{M_i}(s,r) \forall s, r \in S$; **Output**: $\hat{\mathcal{L}}_i$.
$B = S$
$\hat{\mathcal{L}}_i = \emptyset$
while $B \neq \emptyset$

select $s \in B$
define $L_s = \{s\}$
$B = B \setminus \{s\}$
for each $r \in B, r \neq s$
if $d_{M_i}(s,r) < 1$
$L_s = L_s \cup \{r\}$
$B = B \setminus \{r\}$
$\hat{\mathcal{L}}_i = \hat{\mathcal{L}}_i \cup \{L_s\}$

Return: $\hat{\mathcal{L}}_i = \{L_1, L_2, \ldots, L_K\}$

5 Results and Conclusions

For each partition of $\{1,2,3,4,5,6,7,8\}$ we run the algorithm 1 over each part of the partition, obtaining a BIC value. Table 1 gives the numbers corresponding to each language. Table 2 shows the 5 partitions of $\{1,2,3,4,5,6,7,8\}$ with the largest BIC values. On the first line of table 2 we can see in bold face the winning partition which correspond to:

Table 1. Numbering of the languages

Language	Catalan	Dutch	English	Spanish	French	Italian	Japanese	Polish
Number	1	2	3	4	5	6	7	8

Table 2. The five partitions with the largest BIC values

Partition ($\mathcal{M}$)	BIC
{1, 4}, {2, 3, 8}, {5, 6, 7}	**-347488.098872141**
{1}, {2.3.8}, {4}, {5.6.7}	-347509.127775736
{1.4}, {2.3.5.8}, {6.7}	-347517.276944946
{1.4}, {2.8}, {3.5.6.7}	-347519.546013801
{1.4}, {2.8}, {3}, {5.6.7}	-347523.086121394

{Catalan, Spanish}
{Duth, English, Polish}
{French, Italian, Japanese}.

The only discrepancy with the linguistic conjecture on the winning partition of languages is the placement of Japanese which is the only moraic language in the sample and should be alone. We note that the method proposed in García and González-López [10] is able to capture the singularities of Japanese but it has other weaknesses. This misplacement of Japanese also happened on Cuesta-Albertos *et al.* [2] and García *et al.* [11]-García *et al.* [13]. In contrast, the algorithm 1 was particularly efficient in two controversial cases: Polish and Catalan. The first language was included in the stress-timed part while the second language was reported as being similar to Spanish. Despite Polish shows a high syllable complexity, but without the expected vowel reduction for a stress-timed language and Catalan has the same syllabic system as Spanish, although it has some vowel reduction. This suggests that Catalan is not rhythmically different from Spanish.

6 ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by FAEPEX-PRP of University of Campinas, also we would like to thank Franck Ramus for providing the 1648 sentences dataset used on this study. We gratefully acknowledge the support for this research provided by (a) USP project: Mathematics, computation, language and the brain (b) Portuguese in time and space: linguistic contact, grammars in competition and parametric change, FAPESP's project, grant 2012/06078-9 and (c) FAPESP Center for Neuromathematics (grant 2013/ 07699-0, S. Paulo Research Foundation).

References

1. D. Abercrombie. *Elements of general phonetics*, Chicago: Aldine (Chapter 5), 1967.

2. J. Cuesta-Albertos, R. Fraiman, A. Galves, J. Garcia and M. Svarc. Identifying rhythmic classes of languages using their sonority: a Kolmogorov-Smirnov approach. *Journal of Applied Statistics* 34(6), 749-761, 2007.
3. R.M. Dauer. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11, 51-62, 1983.
4. S. Frota, C. Galves, M. Vigário, V. Gonzalez-Lopez and B. Abaurre. The phonology of rhythm from Classical to Modern Portuguese. *Journal of Historical Linguistics* 2(2), 173-207, 2012.
5. A. Galves, J. Garcia, D. Duarte and C. Galves. Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002, International Conference,* 2002.
6. A. Galves, C. Galves, J.E. Garcia, N.L. Garcia and F. Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *The Annals of Applied Statistics* 6(1), 186-209, 2012.
7. J. Garcia, U. Gut and A. Galves. Vocale-a semi-automatic annotation tool for prosodic research. In *Speech Prosody 2002, International Conference,* 2002.
8. J. Garcia and V.A. Gonzalez-Lopez. Minimal markov models. arXiv preprint arXiv:1002.0729, 2010.
9. J.E. García and V.A. González-López. Minimal Markov Models. In *Proceedings of the Fourth Workshop on Information Theoretic Methods in Science and Engineering. Helsinki.* v.1. p.25 - 28, 2011.
10. J.E. García and V.A. González-López. Modeling of acoustic signal energies with a generalized Frank copula. A linguistic conjecture is reviewed. *Communications in Statistics-Theory and Methods* 43(10-12), 2034-2044, 2014.
11. J.E. García, V.A. González-López and M.L.L. Viola. Robust model selection and the statistical classification of languages. In *XI BRAZILIAN MEETING ON BAYESIAN STATISTICS: EBEB 2012,* vol. 1490, no. 1, pp. 160-170. AIP Publishing, 2012.
12. J.E. García, V.A. González-López and R.B. Nelsen. A new index to measure positive dependence in trivariate distributions. *Journal of Multivariate Analysis* 115, 481-495, 2013.
13. J.E. García, V.A. González-López and M.L.L. Viola. Robust Model Selection for Stochastic Processes. *Communications in Statistics-Theory and Methods* 43(10-12), 2516-2526, 2014.
14. F. Ramus, M. Nespor and J. Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265-292, 1999.

Markov Partition Models for Epstein Barr Virus

Jesús E. García[1] and V.A. González-López[2]

[1] Department of Statistics, University of Campinas, R. Sérgio Buarque de Holanda, 651, Campinas (CEP 13083-859), Brazil.
(E-mail: jg@ime.unicamp.br)
[2] Department of Statistics, University of Campinas, R. Sérgio Buarque de Holanda, 651, Campinas (CEP 13083-859), Brazil.
(E-mail: veronica@ime.unicamp.br)

Abstract. In this paper, we introduce a new genome modeling methodology (Minimal Markov Models), which is based on the identification of elements in Markov chain state space that have the same transition probabilities. Thus, the state space is divided into parts and elements in the same part of the partition activate the same random mechanism to select the next element in the sequence. We used the methodology to investigate for differences and similarities between five DNA sequences corresponding to four complete, registered Epstein Barr Virus (EBV) sequences (B95-8-type I, GD1-type I, GD2-type 1, and AG876-type II) and a new EBV strain type I sequence reported in 2012, HKNPC1. From the Minimal Markov Models fitted for each sequence, we found that the sequences GD2 and HKNPC1 (nasopharyngeal carcinoma strains from epithelial cells) were closer to each other than the three other sequences. Our results are consistent with previously investigated aspects in McGeoch and Gatherer [1] and Kwok *et al.* [2].
Keywords: Minimal Markov models, Estimation in Markov chains, Entropy.

1 Introduction

An important issue in the medical literature over the last 35 years is the link between the presence of viruses and cancer diagnosis - Stebbing and Bower [3]. Epstein Barr virus (EBV), a causative agent of infectious mononucleosis, was identified in 1964 in a cultured African Burkitt's lymphoma cell line - Hill [4]. This led to the recognition that EBV is implicated in various types of cancer, such as Burkitt's lymphoma and nasopharyngeal carcinoma. DNA sequence analysis has led to significant advances in understanding and interpreting patterns in DNA sequences that reveal relationships between viruses and cancer. In 1984, the first complete genome sequence of EBV, from a type-1 strain named B95-8, was reported - Baer *et al.* [5]. This sequence was extracted from a North American case of infectious mononucleosis. It is considered as the reference complete sequence of EBV and has since played a central role in EBV research. The last revisited version of this sequence has the accession number

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

NC_007605 and contains the Raji strain sequence. In 2005, the second complete genome sequence of EBV (named GD1), a type I strain described in Zeng *et al.* [6] was obtained. It was isolated from the saliva of a nasopharyngeal carcinoma patient from the province Guangdong in southern China (accession number AY961628). The first complete EBV type II genome sequence, named AG876, was reported in Dolan *et al.* [7] and was described as a Ghanaian case of Burkitt's lymphoma (accession number DQ279927). In 2011, the complete genome sequence of another EBV strain (named GD2) was obtained - Liu *et al.* [8]. The type I strain GD2 was derived from the tumor of a nasopharyngeal carcinoma patient from the Guangdong province in southern China (accession number HQ020558). The genome sequence of EBV, HKNPC1 was reported in 2012 - Kwok *et al.* [2] (accession number JQ009376). It is a type I EBV isolated from a primary nasopharyngeal carcinoma of a Chinese patient in Hong Kong. EBV-type I and EBV-type II are found in all human populations; some studies have shown significant differences between the protein sequences of these strains - Sample *et al.* [9].

Considerable effort had been directed toward comparing the complete sequences of these strains, with the aim of uncovering important distinguishing features, such as the presence/absence of specific genes, see McGeoch and Gatherer [1]. The authors of this report described a genome-wide comparison of these sequences based on single nucleotide polymorphismos (SNPs). The EBV genome sequences B95-8-type I, AG876-type II and GD1-type I were aligned using CLUSTAL W. Seven regions (haplotypes) were identified to determine the incidences of SNPs. In the leftmost regions, the B95-8-type I and AG876-type II sequences were considered very closely similar, but they were clearly distinct from that of GD1 - type I in regions 1 and 2, and all the sequences showed the same pattern in regions 3, 4 and 5. In region 6, B95-8 - type I had different patterns to those of the other two sequences, and in region 7, AG876 - type II had differences to the other two sequences (see figure 2 from McGeoch and Gatherer [1]). Recently, in Kwok *et al.* [2], a phylogenetic analysis with the four complete sequences, B95-8-type I, AG876-type II, GD1-type I, and GD2-type I, and the sequence HKNPC1-type I, revealed HKNPC1 was more closely related to the Chinese nasopharyngeal carcinoma patient-derived strains GD1 and GD2 (for an illustration, see figure 3, A (whole genome) from Kwok *et al.* [2]).

In this paper, we introduce a new genome modeling methodology based on the identification of natural units. The methodology describes the state space, in which the partition is defined by the condition that members of each part have the same transition probability to the next symbol in the sequence. See García and González-López [10] for a complete explanation of the model family, called Minimal Markov Models and see also García and Fernández [11] to other estimation methods under the framework of these models. The model applied in this paper is a generalization of Variable Length Markov Chains models (VLMC), see Rissanen [13], Buhlmann and Wyner [12], Galves *et al.* [14]. VLMC models have been applied to diverse areas, such as genetics (Buhlmann and Wyner [12]) and linguistics (Galves *et al.* [14], García *et al.* [15]). In García *et al.* [15], a robust model selection algorithm for VLMC models is used for the

statistical classification of languages, and the application of this model in this context has been widely investigated in García *et al.* [16]. In this paper, we offer a more flexible approach for the statistical modeling of genome sequences. We estimate a partition of the set of subsequences, such that subsequences belonging to the same part of the partition can be considered as being synonyms because they choose the next element in the sequence with the same transition probability. The Minimal Markov Model reveals the existence of synonym structures in the genome, in the stochastic sense of the term. In Farcomeni [17], this idea is used to extend hidden Markov models.

2 Materials and Methods

2.1 DNA Dataset

The datasets were obtained from http://www.ncbi.nlm.nih.gov/ (NCBI - National Center for Biotechnology Information). The five DNA sequences were the four complete EBV sequences registered so far: (i)B95-8-type I, the reference sequence (accession number NC_007605), see Baer *et al.* [5], named "EBV.WT" according to Kwok *et al.* [2]; (ii)GD1-type I (accession number AY961628) denoted by "GD1", see Zeng *et al.* [6]; (iii)AG876-type II (accession number DQ279927) denoted by "AG876", see Dolan *et al.* [7]; (iv) GD2-type 1 (accession number HQ020558) denoted by "GD2", see Liu *et al.* [8] and (v)the new sequence reported in 2012, an EBV strain type I denoted by HKNPC1 (accession number JQ009376), see Kwok *et al.* [2].

2.2 Minimal Markov Model

Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. Let us call $\mathcal{S} = A^M$ the state space. Denote the string $a_m a_{m+1} \dots a_n$ by a_m^n, where $a_i \in A$, $m \leq i \leq n$. For each $a \in A$ and $s \in \mathcal{S}$, $P(a|s) = \text{Prob}(X_t = a|X_{t-M}^{t-1} = s)$. Let $\mathcal{L} = \{L_1, L_2, \dots, L_K\}$ be a partition of $\mathcal{S}$, $P(L, a) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s, X_t = a),\ a \in A,\ L \in \mathcal{L}$; $P(L) = \sum_{s \in L} \text{Prob}(X_{t-M}^{t-1} = s),\ L \in \mathcal{L}$. If $P(L) > 0, \forall a \in A$, we define $P(a|L) = \frac{P(L,a)}{P(L)}$. We define the statistical model through the following equivalence relation.

Definition 1. Let (X_t) be a discrete time order M Markov chain on a finite alphabet A. We will say that $s, r \in \mathcal{S}$ are equivalent (denoted by $s \sim_p r$) if $P(a|s) = P(a|r)\ \forall a \in A$. For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.

Technically, the equivalence relationship defines a partition of $\mathcal{S}$. The parts of this partition are the equivalence class, i.e. $s, r \in \mathcal{S}$ belongs to different parts if, and only if, they have different transition probabilities. We can interpret that each element of $\mathcal{S}$ on the same equivalence class activates the same random mechanism to choose the next element in the Markov chain. We define now the Markov chain with partition $\mathcal{L}$.

Definition 2. Let (X_t) be a discrete time order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$. We will say that (X_t) is a Markov chain with partition $\mathcal{L}$ if this partition is the one defined by the equivalence relationship $\sim_p$ introduced by definition 1.

The set of parameters for a Markov chain over the alphabet $A = \{a_1, a_2, ..., a_{|A|}\}$ with partition $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ can be denoted by $\{P(a_i|L_j) : 1 \leq i < |A|, 1 \leq j \leq K\}$. If we know the equivalence relationship for a given Markov chain, then we need $(|A| - 1)$ transition probabilities for each part to specify the model. Then, the number of parameters for the model is $|\mathcal{L}|(|A|-1)$, where $|A|$ and $|\mathcal{L}|$ denote the cardinal of A and $\mathcal{L}$ respectively.

Definition 3. Let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$. $L \in \mathcal{L}$ is a good part of $\mathcal{L}$ if $\forall s, s' \in L$

$$\text{Prob}(X_t = . \,|X_{t-M}^{t-1} = s) = \text{Prob}(X_t = . \,|X_{t-M}^{t-1} = s').$$

Definition 4. A partition $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ of $\mathcal{S}$ is a good partition of $\mathcal{S}$ if for each $i \in \{1, \ldots, K\}$, L_i, check definition 3.

The next section shows how we estimate the partition given by definition 2, which we will refer to as "minimal good partition".

Partition Estimation Let x_1^n be a sample of the process $\big(X_t\big)$, $s \in \mathcal{S}$, $a \in A$ and $n > M$. We denote by $N_n(s, a)$ the number of occurrences of the string s followed by a in the sample x_1^n, $N_n(s, a) = \big|\{t : M < t \leq n, x_{t-M}^{t-1} = s, x_t = a\}\big|$. The number of occurrences of s in the sample x_1^n is denoted by $N_n(s)$ and $N_n(s) = \big|\{t : M < t \leq n, x_{t-M}^{t-1} = s\}\big|$. The number of occurrences of elements into L followed by a is given by, $N_n^{\mathcal{L}}(L, a) = \sum_{s \in L} N_n(s, a),\ L \in \mathcal{L}$; the accumulated number of $N_n(s)$ for s in L is denoted by, $N_n^{\mathcal{L}}(L) = \sum_{s \in L} N_n(s),\ L \in \mathcal{L}$.

Definition 5. Let $\mathcal{L}^{ij}$ denote the partition

$$\mathcal{L}^{ij} = \{L_1, \ldots, L_{i-1}, L_{ij}, L_{i+1}, \ldots, L_{j-1}, L_{j+1}, \ldots, L_K\},$$

where $\mathcal{L} = \{L_1, \ldots, L_K\}$ is a partition of $\mathcal{S}$, and for $1 \leq i < j \leq K$ with $L_{ij} = L_i \cup L_j$.

For $a \in A$ we write, $N_n^{\mathcal{L}^{ij}}(L_{ij}, a) = N_n^{\mathcal{L}}(L_i, a) + N_n^{\mathcal{L}}(L_j, a)$; $N_n^{\mathcal{L}^{ij}}(L_{ij}) = N_n^{\mathcal{L}}(L_i) + N_n^{\mathcal{L}}(L_j)$.

Definition 6. Let (X_t) be a Markov chain of order M, with finite alphabet A and state space $\mathcal{S} = A^M$, x_1^n a sample of the process and let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a good partition of $\mathcal{S}$,

$$\begin{aligned} d_{\mathcal{L}}^n(i, j) = \frac{2}{(|A| - 1)\ln(n)} \sum_{a \in A} \Bigg\{ & N_n^{\mathcal{L}}(L_i, a) \ln\left(\frac{N_n^{\mathcal{L}}(L_i, a)}{N_n^{\mathcal{L}}(L_i)}\right) \\ & + N_n^{\mathcal{L}}(L_j, a) \ln\left(\frac{N_n^{\mathcal{L}}(L_j, a)}{N_n^{\mathcal{L}}(L_j)}\right) \\ & - N_n^{\mathcal{L}^{ij}}(L_{ij}, a) \ln\left(\frac{N_n^{\mathcal{L}^{ij}}(L_{ij}, a)}{N_n^{\mathcal{L}^{ij}}(L_{ij})}\right) \Bigg\}. \end{aligned}$$

Remark 1 *i. As a consequence of Theorem 2.1 proved in García and González-López [10], if (X_t) is a discrete time, order M Markov chain on a finite alphabet A and x_1^n is a sample of the process, then for n large enough, $d_{\mathcal{L}}^n(i,j) < 1$ iff L_i and L_j belong to the same part of the true partition.*

ii. The algorithm introduced in García and González-López [18] (using $d_{\mathcal{L}}^n(i,j)$) returns the true partition for the source, this means that under the assumptions of Theorem 2.1 García and González-López [10], $\hat{\mathcal{L}}_n$ given by the algorithm converges almost surely eventually to $\mathcal{L}$, where $\mathcal{L}$ is the partition of $\mathcal{S}$ defined by the equivalence relationship introduced in definition 2.

2.3 Processing the Data Set

The minimal Markov model for each sequence was obtained by applying the algorithm introduced in García and González-López [18]. The 20 amino acid alphabet plus the stop codon was used with IUPAC notation. The concatenation of amino acids observed in the code was the string of realizations x_1^n. The size of each sequence is shown in table 1.

Table 1. Total number of amino acids of each DNA sequence: EBV.WT (accession number NC_007605), GD1 (accession number AY961628), AG876 (accession number DQ279927), GD2 (accession number HQ020558) and HKNPC (accession number JQ009376)

EBV.WT	GD1	AG876	GD2	HKNPC1
54373	57219	54670	52074	54913

For the incomplete sequence HKNPC1, the occurrences of each string were computed separately from the beginning of each stretch of sequence.
The distances between the sequences were obtained using the symmetrized relative entropy, see the next definition.

Definition 7. Given two sequences i and j, let $\hat{Q}_i$ and $\hat{Q}_j$ be the respective models fitted using the model selection algorithm in García and González-López [18]. The symmetrized relative entropy between the sequences i and j is defined by

$$\overline{SRE}_{(i,j)} = \frac{D\left(\hat{Q}_i||\hat{Q}_j\right) + D\left(\hat{Q}_j||\hat{Q}_i\right)}{2},$$

where $D\left(\hat{Q}_i||\hat{Q}_j\right) = \sum_{x\in\chi} \hat{Q}_i(x) \log\left(\frac{\hat{Q}_i(x)}{\hat{Q}_j(x)}\right)$.

3 Results

Table 2 shows the minimal good partition for each DNA sequence. We show each part L, a member of $\mathcal{L}$, as a collection of amino acids. For instance, if L is composed of amino acids F, I, N and Y, then the part L, denoted as the

set $L = \{\text{F, I, N, Y}\}$ means that F, I, N and Y have the same probability of choosing the next symbol in the DNA sequence. More precisely, for any element a in the alphabet of amino acids $A, P(a|\text{ F}) = P(a|\text{I}) = P(a|\text{N}) = P(a|\text{Y})$ and according to definition 1, F, I, N and Y are equivalent.

3.1 The HKNPC1 model is closer to that of GD2 than to that of GD1, EBV.WT, or AG876

Table 2. Minimal Good Partition for each sequence

EBV.WT	GD1	AG876
{stop,K,M,T,V}	{stop,D,H,K,L,M,V}	{stop,E,K,Q}
{A,C,D,H,L,S}	{A,G,P,R,W}	{A,G,P,R,W}
{E,Q}	{C,F,I,N,Y}	{C,D,H,L,M,S,T,V}
{F,I,N,Y}	{E,Q}	{F,I,N,Y}
{G,P}	{S,T}	
{R,W}		
GD2	**HKNPC1**	
{stop,E,K,Q}	{stop,E,K,Q}	
{A,D,G,H,L,M,S,T,V,W}	{A,D,L,M,S,T,V,W}	
{C,F,I,N,Y}	{C,F,H,I,N,Y}	
{P,R}	{G,P,R}	

We note that the minimal good partitions of two nasopharyngeal carcinoma-related EBV strains, GD2 and HKNPC1, are very similar, except for the positions of the amino acids G and H. This result agrees with the findings in Kwok *et al.* [2], in relation to these two sequences. In Kwok *et al.* [2] the authors show that HKNPC1 has closer phylogenetic relationship to GD1 and GD2 than EBV.WT and AG876. We emphasize that both sequences, GD2 and HKNPC1, were obtained from epithelial cells. GD1 (the remaining nasopharyngeal carcinoma EBV strain) was not directly harvested from epithelial tissue, but from saliva. To confirm the proximity between GD2 and HKNPC1, we built a dendrogram using symmetrized relative entropy, defined in 7 (see also figure 1).

Table 3. Symmetrized Relative Entropy

	EBV.WT	GD1	AG876	GD2	HKNPC1
EBV.WT	0.0000	0.0154	0.0130	0.0175	0.0143
GD1	-	0.0000	0.0169	0.0226	0.0225
AG876	-	-	0.0000	**0.0088**	**0.0077**
GD2	-	-	-	0.0000	**0.0074**
HKNPC1	-	-	-	-	0.0000

The dendrogram exposes a proximity also between the group {GD2, HKNPC1}

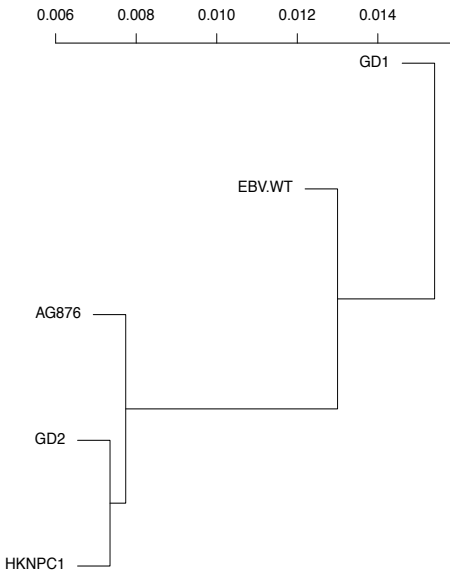

Fig. 1. Dendrogram from the symmetrized relative entropies (see table 3 and definition 7) between the estimated laws for the 5 sequences.

and the strain AG876; this is explained by the low values of the symmetrized relative entropies between GD2 and AG876, and between HKNPC1 and AG876 (see table 3). In addition, if we compare the minimal good partitions of AG876, GD2 and HKNPC1, we can observe the common part {stop,E,K,Q}. Also, we note that for those three sequences, the algorithm detects (i) the equivalence between four amino acids N, Y, F and I; (ii) the equivalence between the six amino acids D, L, M, S, T and V and (iii) the equivalence between the pair of amino acids A and W, and P and R.
The distances between the three strings: EBV.WT, AG876 and GD1, are consistent with the results of McGeoch and Gatherer [1]. See figure 2 in McGeoch and Gatherer [1], HR2 and HR5 cases. In McGeoch and Gatherer [1] the authors use EBV haplotype regions to quantified the proximity between EBV.WT, AG876 and GD1, GD1 is the farthest.

4 Conclusion

The labeling of amino acids that are considered equivalent for genome sequence construction permits the elucidation of the intrinsic stochastic structure of genome sequences. The members of some part of a minimal good partition from some sequences can be considered as natural units of genome architecture and can also reveal stochastic proximity between sequences as shown in this paper. We show, using the minimal Markov model constructed for each sequence, that the nasopharyngeal carcinoma-related EBV strains GD2 and HKNPC1 are

closer, according to symmetrized relative entropy. This finding is in accordance with the results of Kwok *et al.* [2]. Also we obtained results consistent with McGeoch and Gatherer [1] in relation to the distance between EBV.WT, AG876 and GD1. The idea behind this model is to find a genetic profile to facilitate future analysis and comparisons.

5 ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by FAEPEX-PRP of University of Campinas. We also gratefully acknowledge the support for this research provided by (a) USP project: Mathematics, computation, language and the brain (b) Portuguese in time and space: linguistic contact, grammars in competition and parametric change, FAPESP's project, grant 2012/06078-9 and (c) FAPESP Center for Neuromathematics (grant 2013/ 07699-0, S. Paulo Research Foundation).

References

1. D.J. McGeoch and D. Gatherer. Lineage structures in the genome sequences of three Epstein-Barr virus strains. *Virology* 359 (1) p. 1, 2007.
2. H. Kwok, A.H. Tong, C.H. Lin, S. Lok, P.J. Farrel, D.L. Kwong and A. K. Chiang. Genomic sequencing and comparative analysis of Epstein-Barr Virus genome isolated from primary nasopharyngeal carcinoma biopsy. *PLoS ONE* 7 (5): e36939, 2012.
3. J. Stebbing and M. Bower. Epstein-Barr virus in Burkitt's lymphoma: the missing link. *The Lancet Oncology* 10 (4) p. 430, 2009.
4. A.B. Hill. The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine* 58 (5) p. 295, 1965.
5. R. Baer, A.T. Bankier, M.D. Biggin, P.L. Deininger, P.J. Farrell, T.J. Gibson, G. Hatfull, G.S. Hudson, S.C. Satchwell, C. Séguin, P.S. Tuffnell and B.G. Barrell, B.G. DNA sequence and expression of the B95-8 Epstein-Barr virus genome, *Nature* 310 (5974) p. 207, 1984.
6. M.S. Zeng, D.J. Li, Q.L. Liu, L.B. Song, M.Z. Li, R.H. Zhang, X.J. Yu, H.M. Wang, I. Emberg and Y.X. Zeng. Genomic sequence analysis of Epstein-Barr Virus strain GD1 from a nasopharyngeal carcinoma patient. *Journal of Virology* 79 (24) p. 15323, 2005.
7. A. Dolan, C. Addison, D. Gatherer, A.J. Davison and D.J. McGeoch. The genome of Epstein-Barr virus type 2 strain AG876. *Virology* 350 (1) p. 164, 2006.
8. P. Liu, X. Fang, Z. Feng, Y.M. Guo, R.J. Peng, T. Liu, Z. Huang, Y. Feng, X. Sun, Z. Xiong, X. Guo, S.S. Pang, B. Wang, X. Lv, F.T. Feng, D.J. Li, L.Z. Chen, Q.S. Feng, W.L. Huang, M.S. Zeng, J.X. Bei, Y. Zhang, Y.X. Zeng. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. *Journal of Virology* 85 (21) p.11291, 2011.
9. J. Sample, L. Young, B. Martin, T. Chatman, E. Kieff, A. Rickinson and E. Kieff. Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. *Journal of Virology* 64 (9) p. 4084, 1990.
10. J. Garcia and V.A. Gonzalez-Lopez. Minimal markov models. arXiv preprint arXiv:1002.0729, 2010.

11. J.E. García and M. Fernández. Copula based model correction for bivariate Bernoulli financial series. In *11TH INTERNATIONAL CONFERENCE OF NUMERICAL ANALYSIS AND APPLIED MATHEMATICS 2013: ICNAAM 2013,* vol. 1558, no. 1, pp. 1487-1490. AIP Publishing, 2013.
12. P. Buhlmann and A.J. Wyner. Variable length Markov chains. *The Annals of Statistics* 27 (2) p. 480, 1999.
13. J. Rissanen. A universal data compression system. *IEEE Transactions on Information Theory* 29 (5) p. 656, 1983.
14. A. Galves, C. Galves, J.E. Garcia, N.L. Garcia and F. Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *The Annals of Applied Statistics* 6(1), 186-209, 2012.
15. J.E. García, V.A. González-López and M.L.L. Viola. Robust model selection and the statistical classification of languages. In *XI BRAZILIAN MEETING ON BAYESIAN STATISTICS: EBEB 2012,* vol. 1490, no. 1, pp. 160-170. AIP Publishing, 2012.
16. J.E. García, V.A. González-López and M.L.L. Viola. Robust Model Selection for Stochastic Processes. *Communications in Statistics-Theory and Methods* 43(10-12), 2516-2526, 2014.
17. A. Farcomeni. Hidden Markov partition models. *Statistics & Probability Letters* 81 (12) p. 1776, 2011.
18. J.E. García and V.A. González-López. Minimal Markov Models. In *Proceedings of the Fourth Workshop on Information Theoretic Methods in Science and Engineering. Helsinki.* v.1. p.25 - 28, 2011.

The Effect of Missing Visits on GEE, a Simulation Study

Julia Geronimi[1,2] and Gilbert Saporta[2]

[1] Institut de de Recherches Internationales SERVIER, 50 rue Carnot 92150 Suresnes (E-mail: geronimi.julia@gmail.com)
[2] Cedric-Cnam, 292 rue Saint Martin 75141 Paris Cedex 03 (E-mail: gilbert.saporta@cnam.fr)

Abstract. Clinical research is often interested in longitudinal follow-up over several visits. All scheduled visits are not carried out and it is not unusual to have a different number of visits by patient. The Generalized Estimating Equations can handle continuous or discrete autocorrelated response. The method allows a different number of visits by patients. The GEE are robust to missing completely at random data, but when the last visits are fewer, the estimator may be biased. We propose a simulation study to investigate the impact of missing visits on the estimators of the model parameters under different missing data patterns. Different types of responses are studied with an exchangeable or autoregressive of order one structure. The number of subjects affected by the missing data and the number of visits removed, vary in order to assess the impact of the missing data. Our simulations show that the estimators obtained by GEE are resistant to a certain rate of missing data. The results are homogeneous regardless to the imposed missing data structure.
Keywords: Longitudinal data, repeated correlated data, correlation, missing data, simulations, Generalized Estimating Equations.

1 Introduction

Clinical follow-up provides information on changing pattern of diseases. This allows for biological measurements and clinical criterion observation over several visits. Therefore, it is possible to study the link between several potential biological covariates and a clinical response on repeated measurements.

However, observations from the same patient cannot be handled as independent and the correlation among visits must be taken into account. Two of the most common methods which are able to deal with longitudinal data are the Generalized Linear Mixed Model, GLMM as describe by McCulloch [6] and the Generalized Estimating Equations, GEE from Liang and Zeger [5].

GLMM are a subject specific method which introduces a random effect per patient to take into account the longitudinal aspect of observations. Unfortunately, the integration over these random effects distribution may be numerically untractable. GEE are a population specific method which consider the

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

intra-subject correlations by imposing a correlation structure to the response. Advantage of the GEE method is that only correct specification of marginal means is needed for having a consistent and asymptotically normal parameter estimator. We will use this method in this paper. For a discussion on GEE, GLMM and relation between marginal and mixed effect models, reader can refer to the work of Park [9], Heagerty and Zeger[3] and Nelder and Lee[7].

Studies'design provides for a number of visits per patient which is regrettably not always complied. In the case of intermittent missing data this results in blank lines in observation matrix. No classical parametric imputation shall be performed since no information is collected at this date. Moreover the interpolation of these values is difficult because there are often few widely spaced visits which means the prediction is blurred.

Missing data, as defined by Rubin [15], are divided into three categories :

- Missing Completely at Random, like a visit randomly deleted by loss record
- Missing At Random, as a missed visit linked to the length of the study
- Missing Not At Random, such as non presence of a patient related to the latent seriousness of his condition

The GEE estimator is robust to the first case but biased in the other two as explained by Liang and Zeger[5] and Robins *et al.*[13]. In case of dropouts Robins *et al.*[13] introduced an inverse probability of censoring weighted GEE which have been studied by Preisser *et al.*[10]. They proposed a modified version of GEE in which observations or person-visits have weights inversely proportional to their probability of being observed, which is unfortunately not suitable here.

Within this context questions may arise :

- How much the GEE estimator is robust to missing visits?
- Which bias should we consider in case of MAR data?

We provide a simulation study to measure the impact of different missing data patterns on GEE estimators. Second part of this paper gives the GEE approach outline. Simulations plan and their results are shown in section 3 and 4. The paper ends by a conclusion in section 5.

2 Generalized Estimating Equation

When the population-average effect is of interest, the marginal model is commonly used to analyzing longitudinal data. Liang and Zeger[5] proposed the Generalized Estimating Equations to estimate the regression parameter, by only specifying the marginal distribution of the outcome variables in the marginal model. Both continuous and binary responses can be modeled.

Let y_{it}, of expectation μ_{it}, be the response of interest for the subject i at the visit t for $i \in \{1, .., K\}$ and $t \in \{1, .., n_i\}$. Each subject has a set of p measured covariates at each time t denoted x_{it}. For a known function $V(.)$ and a given mean-link function $g(.)$ we have :

$$\mathbb{V}ar(y_{it}) = \phi V(\mu_{it}) \tag{1}$$

$$g(\mu_{it}) = x_{it}^t \beta \tag{2}$$

β is the regression parameter to be estimated, ϕ is the dispersion parameter. We will note Y_i, the $n_i \times 1$ independent response vector and X_i, the $n_i \times p$ measured covariates matrix for subject i. Generalized Estimating Equations are defined by :

$$U(\beta) = \sum_{i=1}^{K} D_i^t V_i^{-1} (Y_i - \mu_i) = 0 \tag{3}$$

D_i is the matrix of partial derivatives with $\partial \mu_{it} / \partial \beta_k$ as its (t, k)-th element. V_i is the working covariance matrix defined by :

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2} \tag{4}$$

where $R_i(\alpha)$ is a working correlation matrix completely described by the parameter vector α of size $s \times 1$. A_i is the diagonal matrix with elements equal to the variance terms $V(\mu_{it})$. If $R_i(\alpha)$ is the true correlation matrix of Y_i then V_i is the true covariance matrix.

Liand and Zeger[5] propose an iterative estimation method. A consistent method (as the moments method) is used to estimate the couple (α, ϕ) for fixed values of $\hat{\beta}$. Then equation (3) is used to estimate $\hat{\beta}$ for fixed values of $(\hat{\alpha}, \hat{\phi})$. This leads to a consistent estimate of β.

The choice of $R_i(\alpha)$ is important. Classic structures are independent, exchangeable or auto-regressive of order 1. Selection criterion for the choice of the working correlation matrix are useful. We quote here just a few : the Quasi-log-likelihood under the independence model Information Criteria from Pan [8], the Correlation Information Criteria from Hin and Wang[4] and Rotnitzky-Jewell's criterion[14]. In order to simplify, we will suppose the working correlation known and of exchangeable or auto-regressive of order one structure.

3 Simulations plan/structure

Two types of responses are studied, a continuous and a binary outcome. Both cases introduce 4 covariates which have been simulated by a Gaussian distribution with an auto-regressive of order one with parameter $\rho = 0.3$. We denote Σ this correlation structure.

3.1 Gaussian response

The response Y_i is a multivariate normal vector with intra-subject correlation structure $R_i(\alpha)$ following the model :

$$Y_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{5}$$

where $x^l \sim \mathcal{N}(0, \Sigma)$ for $l \in \{2, ..., 5\}$. The error vector $\boldsymbol{\epsilon}_i$ is a multivariate normal vector with mean zero and variance matrix $\sigma^2 R_i(\alpha)$. The mean parameter vector is imposed equal to $\beta = (1, 0.5, -0.2, 1, -1)$, where the first component is the intercept. The variance parameter σ^2 is chosen for having a signal/noise ratio of 0.5 as described by Fu[1].

$$\frac{V(x_{it}^t\beta)}{\sigma^2} = \frac{1}{2} \Leftrightarrow \sigma^2 = 2\sum_{l=2}^{5}\beta_l^2 = 4.58 \tag{6}$$

3.2 Binary response

To simulate a binary response, the *logit* link is used and an intra-subject correlation structure equal to $R_i(\alpha)$ is imposed thanks to Qaquish[11].

$$logit(\mathbb{E}(y_{it})) = x_{it}^t\beta \tag{7}$$

where $x^l \sim \mathcal{N}(0, \Sigma)$ for $l \in \{2, ..., 5\}$. The mean parameter vector is imposed equal to $\beta = (1, 0.5, -0.2, 0.3, -0.4)$. The first component is the intercept.

For both kinds of data, the parameters vary as follows according to a full factorial design.

- K, the number of subjects on $\mathcal{K} = \{50, 100, 200, 300\}$
- n, the number of scheduled visits on $\mathcal{N} = \{4, 6, 9\}$
- $R_i(\alpha)$, the correlation structure is either exchangeable or auto-regressive of order one (both admit a scalar $\alpha \to s = 1$)
- α, the unique parameter of correlation on $\mathcal{A} = \{0.1, 0.3, 0.5, 0.6\}$

We simulated 1000 samples that we will called *completed* for each of these 96 scenarios. All of the subjects in these samples get the same number of visits. In order to evaluate the effect of missing visits on the GEE estimators we simulated 1000 other samples that we will called *uncompleted* ou *unbalanced* where we deleted some of the visits on some subjects. The percentage of concerned subjects varies according to $\mathcal{P} = \{10\%, 20\%, 30\%, 50\%\}$ and the number of deleted visits varies according to $\mathcal{V} = \{1, 2, 3\}$.

With the aim of evaluating how robust the GEE estimator is in MCAR and MAR situations, we imposed two different schemes of visits removal. First, we consider a scheme where visits follow a uniform distribution. In that case we can speak of MCAR data. In a second time we consider a probability of

deletion that will increase with the follow-up (i.e. with the number of visits). Last case imposed MAR data. We will talk about uniform unbalanced and increasing unbalanced respectively. All computations are performed using R [12] and GEE fitting performed by the package `geepack` of Halekoh *et al.*[2].

4 Results

A useful criterion for assessing the goodness of an estimator $\hat{\boldsymbol{\theta}}$ is the Absolute Relative Bias defined by $ARB(\hat{\boldsymbol{\theta}}) = \frac{||\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}||}{||\boldsymbol{\theta}||}$. We estimate this criterion by :

$$\widehat{ARB}(\hat{\boldsymbol{\theta}}) = \frac{1}{1000} \sum_{b=1}^{1000} \frac{||\hat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}||}{||\boldsymbol{\theta}||} \tag{8}$$

where $||.||$ is the euclidean norm which boils down to the absolute value when the parameter is a scalar. $\hat{\boldsymbol{\theta}}_b$ is the estimate of $\boldsymbol{\theta}$ on the b-th sample. The mean of the absolute relative gap between the estimator and its target is thus estimated on 1000 samples.

4.1 Continuous response results

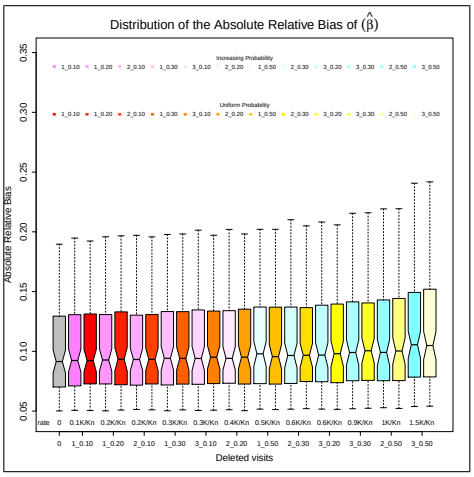

Fig. 1. $\hat{\beta}$ ARB evolution by missing rate for 96 models with a continuous response

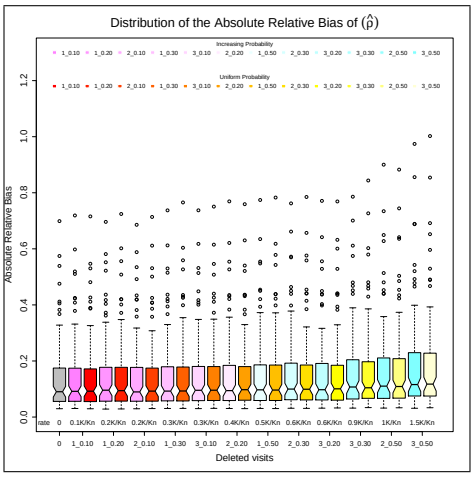

Fig. 2. $\hat{\rho}$ ARB evolution by missing rate for 96 models with a continuous response

Figures 1 and 2 show the distribution of the ARB for the GEE estimator of the parameter β and ρ on the 96 tested models for a continuous response. These graphs compare the two deletion schemes : uniform and increasing.
The boxplots show no differences between the two deletion schemes. Precisely, the difference is between $[-0.005, 0.005]$ for the Absolute Relative Bias of $\hat{\beta}$ and between $[-0.06, 0.06]$ for the ABR of $\hat{\rho}$.

The ARB slightly increases with the missing rate. The median ARB switches from 0.091 to 0.101 for $\hat{\beta}$ and from 0.09 to 0.117 for $\hat{\rho}$.
More precisely, graphics 3, 4 and 5 present the evolution of the Absolute Relative Bias for $\hat{\beta}$ in the case $K = 100$ and $n \in \{4, 6, 9\}$ with increasing unbalanced scheme.

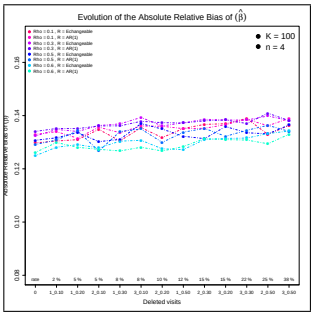

Fig. 3. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 4$ for a continuous response

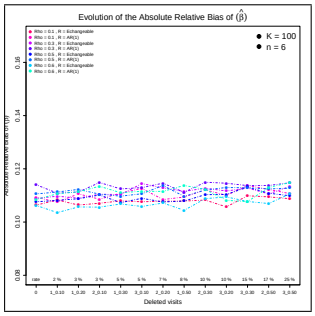

Fig. 4. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 6$ for a continuous response

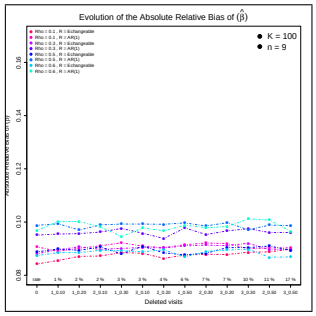

Fig. 5. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 9$ for a continuous response

4.2 Binary response results

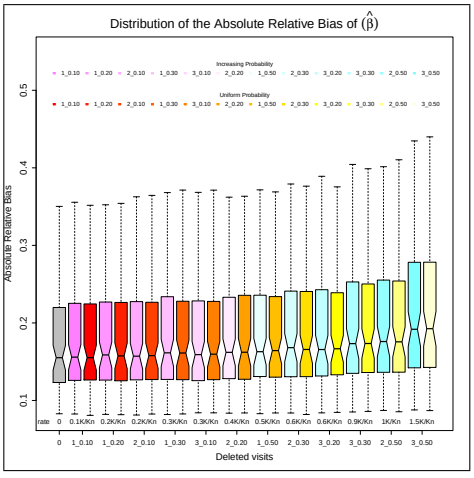

Fig. 6. $\hat{\beta}$ ARB evolution by missing rate for 96 models with a binary response

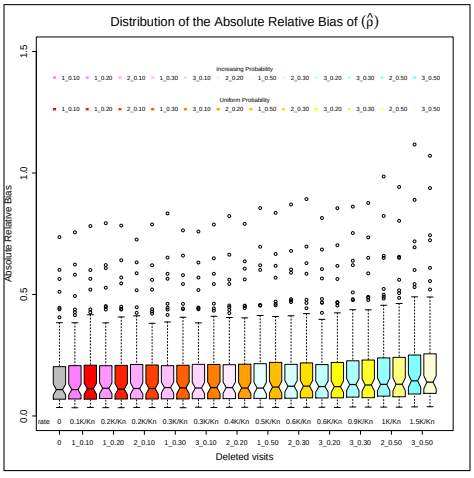

Fig. 7. $\hat{\rho}$ ARB evolution by missing rate for 96 models with a binary response

Graphs 6 and 7 show the distribution of the ARB for the GEE estimator of the parameter β and ρ on the 96 tested models for a binary response. These

graphs compare the two deletion schemes : uniform and increasing.
There are no differences between the two deletion schemes. Some differences in the range of $[-0.005, 0.005]$ and $[-0.015, 0.015]$ have been noted for the Absolute Relative Bias of $\hat{\beta}$ and $\hat{\rho}$ respectively.

The small increase of the ARB is more important for a binary response whith a median ARB switching from 0.155 to 0.193 for $\hat{\beta}$ and from 0.101 to 0.131 for $\hat{\rho}$. Graphs 8, 9 and 10 give more details about the evolution of the Absolute Relative Bias.

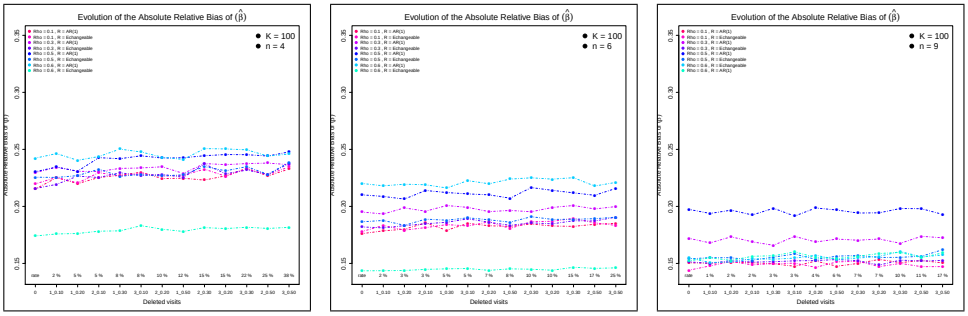

Fig. 8. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 4$ for a binary response

Fig. 9. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 6$ for a binary response

Fig. 10. $\hat{\beta}$ ARB evolution by missing rate for $K = 100$ and $n = 9$ for a binary response

Results on binary response show higher Absolute Relative Bias meaning worst results. Such results were expected since it is more complicated to have an accurate estimator with a binary outcome. Nevertheless both responses, binary and continuous, show the same evolution according to the rate of missing visits. Moreover, both responses point the same lack of differences between uniform unbalanced and increasing unbalanced structure.
Figures 3, 4, 5, 8, 9 and 10 demonstrate how small the increase is with the rate of missing data. The decrease with the number of scheduled visits was expected since it means a lower rate and better estimations.

5 Conclusion

Our simulations show two important issues. First of all, the evolution of the absolute relative bias is similar regardless of the imposed missing data structure. This means that no differences have been highlighted between both schemes. Secondly, the absolute relative bias increases slowly with the missing rate, which means that our imposed missing rate does not disrupt the efficacy of GEE estimator.
We may infer that GEE estimators can be used in studies where MCAR and MAR data are present. Bias induced by MAR is negligible. However, users should pay attention to the missing data scheme and rates used here.

Since it is very complicated to prove the presence of MNAR data, this missing structure has not been studied here. Nevertheless, a complementary study with this type of missing data could bring some more information about expected bias.

References

1. W. Fu. Penalized estimating equations. *Biometrics*, 59:126–132, 2003.
2. U. Halekoh, S. Hojsgaard, and J. Yan. The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006.
3. P. Heagerty and S. Zeger. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26, 2000.
4. L.-Y. Hin and Y.-G. Wang. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28(4):642–658, 2009.
5. K.-Y. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 38:13–22, 1986.
6. C. McCulloch and J. Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.
7. J. Nelder and Y. Lee. Conditional and marginal models: another view. *Statistical Science*, 19(2):219–238, 2004.
8. W. Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125, 2001.
9. T. Park. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, 12(18):1723–1732, 1993.
10. J. Preisser, K. Lohman, and P. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21(20):3035–3054, 2002.
11. F. Qaqish. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90 (2):455–463, 2003.
12. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.
13. J. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.
14. A. Rotnitzky and N. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77 (3):485–497, 1990.
15. D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Modelling patient readmission to hospital using a conditional Markov approach

Andrew S. Gordon, Adele H. Marshall, and Karen J. Cairns

Centre for Statistical Sciences and Operational Research (CenSSOR), Queen's University Belfast, Northern Ireland, UK
(E-mail: agordon17@qub.ac.uk)

Abstract. With the increase in the length of time people are living, comes an increase in the strain placed upon hospitals due to the rising number of elderly patients requiring specialised care. Unfortunately, this may lead to a compromise in the quality of care that pressurised hospitals can deliver. However, if the recurring nature of elderly patient movements between the community and hospital can be understood, then it may be possible for alternative provisions of care in the community to be put in place for patients before readmission to hospital is required. The *conditional Coxian mixture approach* is presented, taking the form of a mixture of Coxian phase-type distributions incorporating Bayes' Theorem. This method helps to bring about understanding by modelling patient pathway through successive stages of care in the form of an aggregate Markov model, whereby length of stay at each stage is conditioned on the length of stay from the previous stage. For the purpose of demonstration, patient hospital and community data is simulated, providing an illustration of the model applied to a synthetic data set.
Keywords: Bayes' theorem, Coxian phase-type distribution, length of stay, readmission, survival analysis.

1 Introduction

The number of elderly people living in the United Kingdom is rising. In 2015, there are almost 11.8 million people aged 65 years old or more with this number projected to increase to 15.5 million by the year 2030 and to reach 20 million by the year 2050 [1]. Unfortunately, however, long life is no guarantee of good health and as a result, the use of health services increases considerably with age, whereby the majority of resources are required in the final year of life [2]. The net result is that this places a strain on health services and in particular, hospital care, meaning that patients often endure a less satisfactory hospital experience, usually exacerbated by longer waiting times for treatment and care [3]. If the process of elderly hospital admissions can be modelled as part of a network comprising the various pathways that patients may take through care, then further insight, understanding and evidence of both this process and how it can be operated efficiently, could be provided to hospital managers ensuring that hospital resources are used to their full potential, thereby minimising waste. Indeed, better planning of the allocation of hospital beds along with coherent collaboration and organisation with the other care facilities in the network could have a greater impact on both the health and wellbeing of patients. Throughout the last ten years, phase-type distributions [4] and in

16th ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece

particular, the Coxian phase-type distribution [5] have been extensively used to model patient length of stay in hospital [6][7][8]. The methodology presented in this paper is the *conditional Coxian mixture approach* which aims to facilitate hospital managers in understanding the role that patient discharge to the community has on the overall cycle of care. This is achieved through the incorporation of a number of stages (each taking the form of a mixture of Coxian phase-type distributions) representing hospital readmissions and discharges to the community, in the overall aggregate Markov model.

2 Methodology

2.1 The Coxian phase-type distribution

The Coxian phase-type distribution [5] is a special type of phase-type distribution (a random variable which describes the time until absorption of a continuous-time Markov process) in which the system starts in the first state (or phase) with the states having an inherent and well-defined order. It is not permitted for the system to proceed directly from one state to another more than one transition along the sequence, nor is it possible to go backwards to any previous state. Introducing notation, the Coxian phase-type distribution is defined as follows: let $X(t); t \geq 0$ be a Markov chain in continuous time with states $1, 2, ..., m, m+1$, where state $m+1$ is the absorbing state, the rest being transient and ordered. The process, necessarily, starts in state 1 ie. $X(0) = 1$. Figure 1 gives an example of such a Coxian phase-type distribution, with transitions occurring in a small time interval h. In the diagram, λ_i represents the transition rate from transient state i to the next transient state $i+1$ and μ_i represents the transition rate from transient state i to the absorbing state, $m+1$.

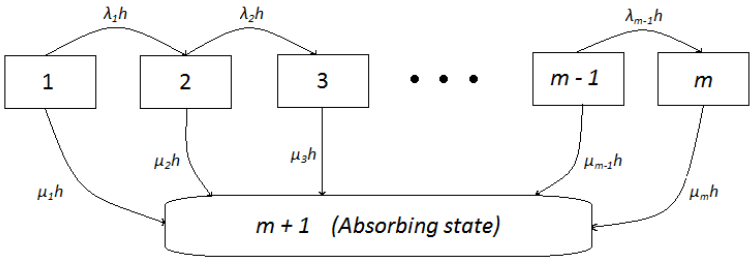

Fig. 1. Example of a Coxian phase-type distribution with m transient states.

The generator matrix $\mathbf{Q}$, required for determining the probability density function for a general phase-type distribution, may then be calculated through the following equation:

$$\mathbf{Q} = \lim_{h \to 0} [\frac{\mathbf{P}(h) - \mathbf{I}}{h}];$$

where $\mathbf{P}(h)$ is the m x m transition matrix with elements $P_{ij}(h)$ equal to the probability of moving between states i and j in small time h and $\mathbf{I}$ is the m x m identity matrix. The generator matrix $\mathbf{Q}$ for the Coxian phase-type distribution shown in Figure 1 is given by

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_1 + \mu_1) & \lambda_1 & 0 & ... & 0 & 0 \\ 0 & -(\lambda_2 + \mu_2) & \lambda_2 & ... & 0 & 0 \\ 0 & 0 & -(\lambda_3 + \mu_3) & ... & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & ... & -(\lambda_{m-1} + \mu_{m-1}) & \lambda_{m-1} \\ 0 & 0 & 0 & ... & 0 & -\mu_m \end{pmatrix}$$

The vector $\mathbf{q}$ is the m x 1 vector containing the transition rates between each of the transient states and the absorbing state, taking the following form:

$$\mathbf{q} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{m-1} \\ \mu_m \end{pmatrix}$$

The initial probability vector, $\mathbf{p}$, is a 1 x m vector with first element equal to 1 and remaining m - 1 elements equal to zero. This is to reflect the requirement that the system must begin in the first state. As a result, the probability density function for the Coxian phase-type distribution may be calculated using the following expression:

$$f(t) = \mathbf{p}\mathrm{e}^{\mathbf{Q}t}\mathbf{q} \tag{1}$$

The states, quite often referred to as phases, in such a system may be used to describe the stages of a process which terminates at some particular instant. For example, in the case of determining the duration that geriatric patients spend in hospital, transitions through the ordered transient states have been interpreted to correspond to the different stages of care that patients go through whilst at hospital, such as short-stay, rehabilitation and long-stay care; where patients may eventually discharge, transfer or die at any of the states [9].

2.2 Mixture of Coxian phase-type distributions

The method of a mixture of phase-type distributions, proposed by Garg *et al.* [10], allows provision for more than one absorbing state (and indeed, more than one cohort of patient moving through the system), whereby it may now be ascertained directly from the model, the precise rates with which patients move into *each* of the absorbing states. This is particularly useful for modelling the

movement of patients between the hospital and the community, since it is possible to precisely identify which patients have been discharged to the community, rather than using a potentially inaccurate sample of those who have simply undergone 'global' absorption from the hospital stage. This is a disadvantage of using the Coxian phase-type distribution on its own for the particular problem under consideration - when the wider network of possible patient movement is considered, it is important to track precisely the destination of patients on leaving hospital. Mixed distributions have become particularly popular for use in conjunction with the Coxian phase-type distribution, resulting in a *mixture of Coxian phase-type distributions (MC-Ph distribution)* [11]. This enables the modelling of C different cohorts whereby each has a different survival distribution (in turn leading to a different distribution for the length of stay duration) and each cohort may undergo absorption to multiple absorbing states. Using the general form of the probability density function for the Coxian phase-type distribution shown by equation 1, the MC-Ph distribution for C cohorts and M absorbing states has probability density function given by

$$f(x) = \sum_{c=1}^{C} \mathbf{p}_c \mathrm{e}^{\mathbf{Q}_c t} \mathbf{q}_c \tag{2}$$

with the following notation:

$$\mathbf{Q}_c = \begin{pmatrix} -(\lambda_1^c + \sum_j \mu_{1j}^c) & \lambda_1^c & ... & 0 \\ 0 & -(\lambda_2^c + \sum_j \mu_{2j}^c) & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & -\sum_j \mu_{k_c j}^c \end{pmatrix} \tag{3}$$

$$\mathbf{q} = \begin{pmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \vdots \\ \mathbf{q}_C \end{pmatrix} \qquad \text{where} \qquad \mathbf{q}_c = \begin{pmatrix} \mu_{11}^c & \mu_{12}^c & ... & \mu_{1M}^c \\ \mu_{21}^c & \mu_{22}^c & ... & \mu_{2M}^c \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{k_c 1}^c & \mu_{k_c 2}^c & ... & \mu_{k_c M}^c \end{pmatrix} \tag{4}$$

$$\mathbf{p}_c = (\alpha_c, 0, ..., 0)$$

where the c^{th} mixture component is a phase-type distribution with k_c transient states and the absorbing state, and α_c is the probability that the system takes the form of this c^{th} component. It should be noted that the MC-Ph distribution is a special type of Coxian phase-type distribution, where the probability of moving between states corresponding to different cohorts is zero. This means that equations 1 and 2 are equivalent, with $\mathbf{Q}$ equal to a diagonal C x C matrix with diagonal elements equal to $\mathbf{Q}_c$ for $c = 1 : C$ and $\mathbf{p}$ is simply the concatenation of the $\mathbf{p}_c$ vectors. Some further important results may be calculated: the cumulative distribution function for the probability of absorption into a

particular absorbing state and the vector of moment generating functions for the unconditional length of stay in the transient states prior to discharge to each absorbing state. This latter result enables the mean and variance of the length of stay in each transient state to be estimated for different cohorts of patients moving through the system.

2.3 An approach using joint probabilities

In the next section, the proposed methodology is presented. Firstly, however, it is necessary to describe how it is possible to model the patient pathway between successive types of care, as derived previously from work carried out by Xie *et al.* [12] but in the form of a general phase-type distribution. Each type of care is known as a *stage* within the system eg. the first hospital stage or the second community stage. The aggregate model ie. the concatenation of all stages together, permits the system to move from any state in a particular stage to the first state in the subsequent stage. This is to reflect both the discharge of patients from any state in the hospital stage to the community and also the readmission of patients from any state in the community stage to the hospital. There exists a single, global absorbing state representing the death of the patient, which may be reached from any state belonging to any stage. Figure 2 shows a simplification of this process containing two stages A and B.

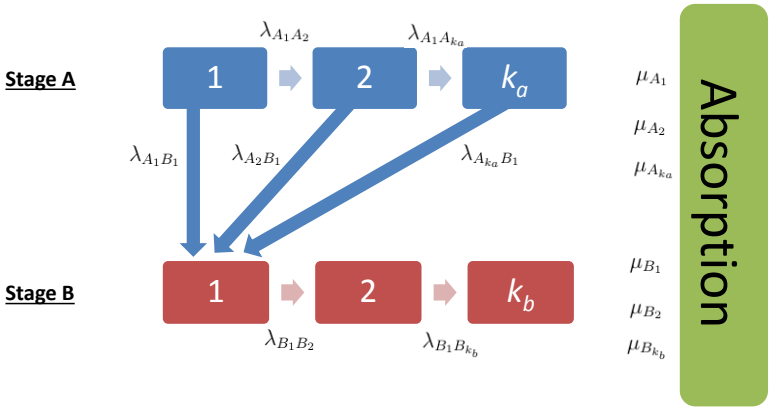

Fig. 2. Representation of each stage as a mixture of Coxians with two absorbing states.

The generator matrix for the aggregate model is as follows:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_A & \mathbf{T}_{AB} \\ \mathbf{0}_{BA} & \mathbf{Q}_B \end{pmatrix} \tag{5}$$

In this formulation, $\mathbf{Q}_A$ is the sub-generator matrix representing stage A. This is cast as follows, where stage B (containing k_b states) is the next stage of care along from stage A (containing k_a states), $\lambda_{A_iA_j}$ is the rate of transition between states i and j in stage A, $\lambda_{A_iB_1}$ is the rate of transition between state i of stage A and the first state of stage B and μ_{A_i} is the rate of absorption from state i of stage A (arrows not shown in Figure 2 to maintain clarity):

$$\mathbf{Q}_A = \begin{pmatrix} -(\lambda_{A_1A_2}+\lambda_{A_1B_1}+\mu_{A_1}) & \lambda_{A_1A_2} & \dots & 0 \\ 0 & -(\lambda_{A_2A_3}+\lambda_{A_2B_1}+\mu_{A_2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -(\lambda_{A_{k_a}B_1}+\mu_{A_{k_a}}) \end{pmatrix}$$

The element of equation 5 on the super-diagonal ie. the trans-stage sub-matrix $\mathbf{T}_{AB}$, represents patients transferring from each state of stage A to the first state of stage B. As a result, $\mathbf{T}_{AB}$ is a sub-matrix of dimension k_a x k_b, where k_a and k_b are the number of states in stage A and stage B, respectively. Due to the fact that patients may only enter the first state of any particular stage from the previous stage, $\mathbf{T}_{AB}$ contains non-zero elements in its first column only, with all other elements equal to zero. An example of the form of $\mathbf{T}_{AB}$ is as follows:

$$\mathbf{T}_{AB} = \begin{pmatrix} \lambda_{A_1B_1} & 0 & \dots & 0 \\ \lambda_{A_2B_1} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{A_{k_a}B_1} & 0 & \dots & 0 \end{pmatrix}$$

Finally, from equation 5, the element $\mathbf{0}_{BA}$ is a zero/null sub-matrix of dimension k_b x k_a. It can be seen that, although still a phase-type distribution, this aggregate model does not have a standard Coxian format, due to the non-zero probability of moving between any state in a particular stage to the first state in the subsequent stage. In fact, the only part of the model retaining strict Coxian format is the sub-generator matrix in the bottom right-hand corner of $\mathbf{Q}$: $\mathbf{Q}_B$, since there are no further stages for the system to progress to, in the example given by equation 5. The probability density function for a patient undergoing absorption during the first stage is simply given by an expression analogous to equation 1, with each component given subscripts pertaining to the initial hospital stage, ie.

$$f_1(t) = \mathbf{p}_A \mathrm{e}^{\mathbf{Q}_A t_A} \mathbf{q}_A$$

The probability density function for a sequence of two stages is derived from the work of Fredkin and Rice [13]. Such a function for a patient moving through a stage A in time t_A and undergoing absorption after time t_B in stage B is given by:

$$f_2(\{t_i\}) = \mathbf{p}_A \mathrm{e}^{\mathbf{Q}_A t_A} \mathbf{T}_{AB} \mathrm{e}^{\mathbf{Q}_B t_B} \mathbf{q}_B \tag{6}$$

Equation 6 is the form that the probability density function takes for a system moving through two successive stages of an overall aggregate Markov process. The advantage of simply extending this model is that patients can be tracked from their initial hospital admission all the way through the stages of care until they either enter the end of life state or the time-frame of the study finishes. Nevertheless, the drawback with extending this approach is that as the number of stages (ie. readmissions) increases, the probability density function values given by equations representing longer concatenations of successive stages become very small, even when a given patient has lengths of stay in each stage which are to be expected. Unfortunately, this would not be as useful for hospital managers managing beds in its current form.

2.4 A conditional approach

The previous methodology achieved success [12] perhaps due to the low number of stages incorporated in the model (two), whereas when the approach is applied to hospital readmissions, the number of stages will be larger, due to the frequency with which elderly patients are readmitted to hospital care. It is, however, possible to use the result given by equation 6 to model the desired process in a different way. Instead of calculating the joint probability of a patient having lengths of stay t_1, t_2 and t_3 in, for example, three successive types of care, a more useful insight into patient movements may be gained if a *conditional approach* is considered whereby the length of stay for a particular type of care is conditioned on the length of stay observed in the most recent type of care. This may be achieved through the use of Bayes' theorem which, given the previous notation, may be cast as follows: let A denote the length of stay for an individual at the previous stage and let B denote the length of stay for the same individual at the current stage, then

$$\mathrm{P}(B = t_2 | A = t_1) = \frac{\mathrm{P}(A = t_1 \cap B = t_2)}{\mathrm{P}(A = t_1)} \tag{7}$$

Noting that $\mathrm{P}(A = t_1 \cap B = t_2)$ is an example of the joint probability for two successive events occurring, then it is permissible to calculate this quantity through the use of equation 6. As a result, the conditional probability becomes:

$$\begin{aligned} \mathrm{P}(B = t_2 | A = t_1) &= \frac{\mathrm{P}(A = t_1 \cap B = t_2)}{\mathrm{P}(A = t_1)} \\ &= \frac{\mathbf{p}_A \mathrm{e}^{\mathbf{Q}_A t_1} \mathbf{T}_{AB} \mathrm{e}^{\mathbf{Q}_B t_2} \mathbf{q}_B}{\mathbf{p}_A \mathrm{e}^{\mathbf{Q}_A t_1} \mathbf{q}_A} \end{aligned} \tag{8}$$

Equation 8 may be used to calculate the probability that a patient experiences a length of stay equal to t_2 in care stage B, given that in the previous stage of care (stage A) they stayed for a time t_1. This may be accomplished by using equation 8 in place of equation 1 when calculating the likelihood function for the model. The *conditional Coxian mixture* model combines this conditional

approach with the theory on mixtures of Coxians described earlier with a view to representing the flow of patients from the initial hospital stage, through their first discharge to the community, through the first readmission and so on. As an illustration of the model, a synthetic data set of patients is simulated for an example scenario of three stages of care: an initial hospital stage, the first discharge to community and finally the first readmission to hospital. At each stage, patients may undergo absorption into one of two absorbing states: the global absorbing state (patient death) and either discharge to community (for hospital stages) or readmission to hospital (for community stages). Figure 3 shows the general m-state flow diagram for a stage of the system with two absorbing states: one representing patient death and the second representing the patient moving on to the next stage of care.

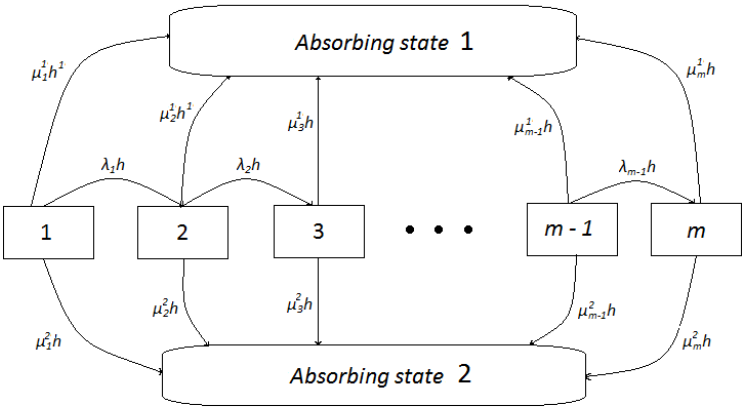

Fig. 3. Representation of each stage as a mixture of Coxians with two absorbing states.

The initial hospital stage is modelled simply by a mixture of Coxian phase-type distributions with one cohort and two absorbing states, without conditioning. The probability density function for this first part of the model is given by equation 1, acknowledging the caveats shown in equations 3 and 4. This may then be fitted for an increasing number of phases using the method of maximum likelihood until the BIC [14]/AIC [15] have been optimised. The corresponding number of phases is then considered optimal for this stage of the model. Once this first stage has been modelled, the focus turns to the second stage (which will be conditioned on the first stage). This (community) stage is once again modelled by a mixture of Coxian phase-type distributions consisting of one cohort and two absorbing states (end of life and readmission to hospital), however the probability density function is given by equation 8. As inputs, this stage of the model requires the optimal parameter estimates from the previous stage

with the exception of the global absorption rates, in addition to the length of stay in the previous stage for each patient. In a similar way to the first stage, this part of the model is fitted for an increasing number of phases until the likelihood function is optimised whereupon the corresponding number of phases is considered optimal. The model, as a whole, is easily extendable to a greater number of stages without risk of the probability density function tending to zero. The previous theory may be replicated for as many readmissions to hospital/discharges to community as is necessary (each time using equation 8 to calculate the probability density function) before the distribution of length of stay between successive readmissions/discharges shows little change. All calculations using this model are computed using *MATLAB* [16].

3 Simulation of data using the proposed model and results of the fitting process

Times may be simulated through the use of the model survivor function $S(t)$. Substitution of a random number from the uniform distribution (0, 1) for $S(t)$ and solving for t results in a simulated time. This process may then be repeated for the desired number of simulated times. In order to solve for t, the Newton Raphson method [17] is utilised due to the unavailability of an analytical solution for the general phase-type distribution. It is the aim to model the pathway of these patients from the initial hospital stage, through the first community stage and finally through the first hospital readmission stage. Since for each stage (aside from the initial stage) the model requires two time vectors (that for the previous stage as well as that to be generated for the current stage) it is important to retain the previous time vector for the generation of the next set of times. An indicator variable is also created, denoting how far through the stages each patient has progressed before entering the global absorbing state. Figure 4 shows the actual pathway between the three stages of care that are to be modelled. The rate parameters displayed in Figure 4 use the following notation: λ_j^i denotes the rate of transition from the j^{th} state to the next state along in the i^{th} stage of the model. Furthermore, μ_j^i denotes the rate of transition rate from the j^{th} state in the i^{th} stage to the absorbing state representing progression to the next stage of the model. Similarly, μ_{aj}^i represents the transition rate from the j^{th} state in the i^{th} stage of the model to the global absorbing state. For ease of presentation in Figure 4 the latter transition rates are not accompanied by arrows. The process which is being modelled consists of three states in hospital, two states in the community and three states in hospital readmission; a total of three stages. Simulation of the data vectors is conducted using *MATLAB*, as is the fitting process.

Length of stay data for 10,000 patients were simulated for the initial hospital stage. Of these, a random sample of 2,500 move to the global absorbing state with the remaining 7,500 virtual patients undergoing discharge to the community stage. From this stage, 3,000 patients are randomly chosen to move to the global absorbing state and the remaining 4,500 undergo readmission to the final hospital stage. At this final stage, 2,500 patients are randomly selected

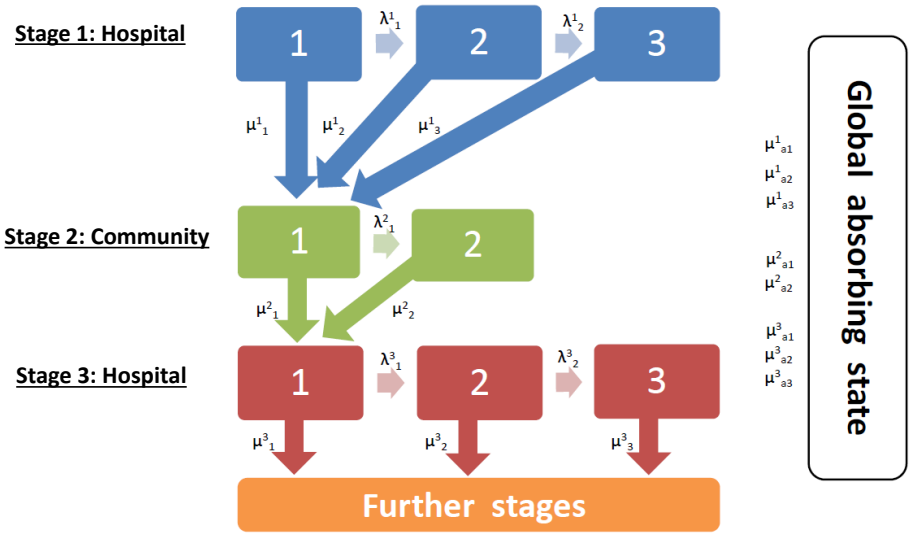

Fig. 4. The succession of stages to be modelled.

Table 1. True values and parameter estimates for the simulated data

Parameter	True value	Estimate	Standard Error	95% Confidence Interval
λ_1^1	0.0800	0.1371	0.1743	[0.0000, 0.4787]
λ_2^1	0.0650	0.0286	0.0145	[0.0002, 0.0570]
μ_1^1	0.0580	0.0623	0.0353	[0.0000, 0.1315]
μ_2^1	0.0450	0.0442	0.0068	[0.0309, 0.0575]
μ_3^1	0.0290	0.0259	0.0030	[0.0200, 0.0318]
μ_{a1}^1	0.0610	0.0628	0.0066	[0.0499, 0.0757]
μ_{a2}^1	0.0540	0.0539	0.0053	[0.0435, 0.0643]
μ_{a3}^1	0.0430	0.0381	0.0037	[0.0308, 0.0454]
λ_1^2	0.0750	0.0574	0.0260	[0.0064, 0.1084]
μ_1^2	0.0470	0.0461	0.0025	[0.0412, 0.0510]
μ_2^2	0.0340	0.0334	0.0015	[0.0305, 0.0363]
μ_{a1}^2	0.0510	0.0500	0.0021	[0.0459, 0.0541]
μ_{a2}^2	0.0380	0.0366	0.0018	[0.0331, 0.0401]
λ_1^3	0.0660	0.0401	0.0605	[0.0000, 0.1587]
λ_2^3	0.0580	0.0014	0.0049	[0.0000, 0.0110]
μ_1^3	0.0490	0.0506	0.0626	[0.0164, 0.0858]
μ_2^3	0.0420	0.0329	0.0363	[0.0402, 0.0562]
μ_3^3	0.0320	0.0177	0.0208	[0.0000, 0.0850]
μ_{a1}^3	0.0530	0.0511	0.0177	[0.0000, 0.1733]
μ_{a2}^3	0.0480	0.0482	0.0041	[0.0000, 0.1040]
μ_{a3}^3	0.0410	0.0231	0.0316	[0.0000, 0.0585]

to move to the next community stage (not modelled), with the remaining patients entering the global absorbing state. This is where the simulation of data in the synthetic study ends. From Table 1 it can be seen that the parameter estimates returned through the model fitting process are close to the true values initially used in the simulation, with all but two of the twenty one true parameters falling within the 95% confidence interval for the corresponding estimate. It should be noted that since the transition rates should be strictly positive to have real-world meaning, the confidence intervals are bounded at the lower end with zero, even if they are calculated as being negative. The average length of stay in each state of a particular stage may be calculated analagously to the method derived by Marshall and McClean [6] which makes use of the calculated parameter estimates. Using the parameter estimates from the above simulated data, the average length of stay in each of the states of the initial hospital stage are calculated as 4, 8 and 16 days, respectively. Upon discharge to the community stage, the two states may represent an initial recovery period during which the discharged patient is perhaps more at risk of being readmitted to hospital, along with a lower risk (or more lengthy) spell in the community. The expected length of stay in each of these latent states is 7 and 14 days, respectively. Finally, when readmitted to hospital for the first time, it is expected that patients will spend 7, 12 and 25 days in each of the short-stay, medium-stay and long-stay states, respectively. These results are consistent with what is expected when geriatric patient care is considered, arising from the implementation of realistic transition rates in the simulation of the patient data.

4 Discussion

Research has been undertaken into the development of a new model called the *conditional Coxian mixture approach*, which is able to represent patient movements between hospital and the community, including readmissions to hospital. The approach extends previous research which has been successful in modelling individual hospital length of stay, but thus far has been less successful in incorporating the additional aforementioned patient movements. The model takes the form of a general phase-type distribution model, specifically a mixture of Coxian phase-type distributions incorporating a Bayesian component, allowing the time until a particular event occurs (eg. death or readmission to hospital) to be conditioned on the time taken to exit the previous stage. A synthetic data set consisting of length of stay variables for 10,000 patients has been simulated over three distinct stages of care: the initial hospital stage, the subsequent discharge to the community and the readmission to hospital. The model has been validated using this synthetic data, with the parameter estimates obtained through the fitting process closely reflecting those used in its simulation. It is hoped that with the further incorporation of patient covariates in model, this may lead to hospital managers having a better understanding of individual patient movement habits. Once ascertained, alternative measures of care may be put in place so that the prospective patient may receive appropriate care in the community, thereby relieving some of the pressure on hospitals.

References

1. Online report (2010). The ageing population. http://www.parliament.uk/business /publications/research/key-issues-for-the-new-parliament/value-for-money-in-public-services/the-ageing-population/ (Accessed 9/1/14).
2. C.R. Victor and I. Higginson. Effectiveness of care for older people: a review, *Quality in Health Care*, **3**, 210–216,1994.
3. Online report (2013). Hospital 'bed blocking' pressure worry for older patients. http://www.bbc.co.uk/news/uk-wales-22275940 (Accessed 10/1/14).
4. M. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*, Courier Dover Publications, 1981.
5. D.R. Cox. A use of complex probabilities in the theory of stochastic processes, *Proc. Camb. Phil. Soc.*, **51**, 313–319, 1955.
6. A.H. Marshall and S.I. McClean. Using Coxian Phase-Type Distributions to Identify Patient Characteristics for Duration of Stay in Hospital, *Health Care Management Science*, **7**, 285–289, 2004.
7. A.H. Marshall, B. Shaw and S.I. McClean. Estimating the costs for a group of geriatric patients using the Coxian phase-type distribution, *Statistics In Medicine*, **26**, 2716–2729, 2006.
8. B. Shaw and A.H. Marshall. Modelling the flow of congestive heart failure patients through a hospital system, *Journal of the Operational Research Society*, **58**, 212–218, 2007.
9. M.J. Faddy. Examples of fitting structured phase-type distributions, *Appliend Stochastic Models and Data Analysis*, **10**, 247–255, 1994.
10. L. Garg, S. McClean, M. Barton, E. El-Darzi and B. Meenan. Clustering patient length of stay using mixtures of Gaussian models and phase-type distributions, *Computer Based Medical Systems (CBMS), 2009 IEEE 22nd International Symposium on.*, 1–7, 2009.
11. S. McClean, L. Garg, M. Barton and K. Fullerton. Using Mixed Phase-type Distributions to Model Patient Pathways, *Computer Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on.*, 172–177, 2010.
12. H. Xie, T.J. Chausselet and P.H. Millard. A continuous-time Markov model for the length of stay of elderly people in institutional long-term care, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1, **168**, 51–61, 2005.
13. D.R. Fredkin and J.A. Rice. On aggregated Markov processes, *Journal of Applied Probability*, **23**, 208–214, 1986.
14. G. Schwarz. Estimating the time dimension of a model, *The Annals of Statistics*, **6**, 461–464, 1978.
15. H. Akaike. A new look at the statistical identification model, *Automatic Control, IEEE Transactions on.*, 6, **19**, 716–723, 1974.
16. *MATLAB ®Reference Guide*, The MathWorks Inc: Natick, Massachussettes, 1992.
17. E. Halley, I. Newton and J. Raphson. *Tracts on the Resolution of Affected Algebraick Equations by Dr. Halley's, Mr. Raphson's and Sir Isaac Newton, Methods of Approximation...*, J. Davis, 1800.

Quality of Tests Defined by Bans

Alexander Grusho[1], Nick Grusho[2], and Elena Timonina[3]

[1] Faculty of Computational Mathematics and Cybernetics, Moscow State University
Leninskie gory, 119991 GSP-1, Moscow, Russia
(E-mail: grusho@yandex.ru)
[2] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Vavilova str., 44/2, 119333, Moscow, Russia
(E-mail: info@itake.ru)
[3] Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Vavilova str., 44/2, 119333, Moscow, Russia
(E-mail: eltimon@yandex.ru)

Abstract. Sets of infinite sequences with elements which belong to finite different alphabets are considered.

In the paper it is offered to use the statistical techniques with probabilities of false alarms equal to zero. This class of statistical decisions is based on concept of bans of probability measures in a finite space. Conditions under which power functions of statistical criteria accept value 1 on a finite step are found. These conditions are formulated in terms of supports of probability measures. If the conditions are fulfilled we present the brute force algorithm to construct tests defined by bans with power function equals to 1.
Keywords: Bans of probability measures in finite spaces, statistical criteria, power function of criteria, monitoring systems.

1 Introduction

Mathematical models for monitoring systems are actual in our days. Suppose that such monitoring uses statistical techniques. Let the trajectories of functioning of such systems be represented by infinite sequences in which each coordinate accepts value in the finite different fixed alphabet.

Application of statistical techniques on a set of infinite sequences demands a probability measure P which describes the correct behavior of analyzable system. The wrong system behavior is described by a probability distribution Q. Different wrong behaviors of the technological system can be described by different distributions of probabilities on space of the infinite sequences.

In practice the monitoring system of a process observes initial sections of trajectories and for chosen step n it tests the hypothesis $H_{0,\,n}$ that the distribution of the observed section of trajectory is defined by probability distribution measure P_n which is the projection of measure P on the first n coordinates.

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

The alternative hypothesis $H_{1,n}$ in the elementary case is defined by measure Q_n which is projection of measure Q on the first n coordinates. Criteria of testing of hypotheses $H_{0,n}$ against alternatives $H_{1,n}$ allow to make the decision about the wrong behavior of a system.

The basic problem for developers of such monitoring systems is the false alarms appearance when the correct behavior of technological process is perceived as wrong Axelson[1]. False alarms lead to necessity of the manual analysis of the reasons of the wrong system behavior.

In the paper it is offered to use the statistical techniques for monitoring with probabilities of false alarms equal to zero. This class of statistical decisions is based on concept of the ban Grusho and Timonina[4] , Grusho *et al.*[3]. The ban of a probability measure in the considered scheme is a vector for which probability of its appearance is equal to 0 in a finite projection of measure.

Any statistical criterion for testing $H_{0,n}$ against $H_{1,n}$ is defined by a critical set S_n of vectors of length n. When the observed vector is in S_n then it leads to the acceptance of alternative $H_{1,n}$. If all vectors in S_n are bans of a measure P_n, say that the criterion is defined by bans of a measure P.

Existence and some properties of the criteria defined by bans were researched in papers Grusho and Timonina[4], Grusho *et al.*[3], [5]. Conditions of consistency of sequence of the statistical criteria defined by bans have been found.

In this paper conditions under which power functions of criteria accept value 1 on a finite step are found. These conditions are formulated in terms of supports of probability measures for the main measure P on space of the infinite sequences and for alternatives. If the conditions are fulfilled we present the brute force algorithm to construct tests defined by bans with power function equals to 1.

The article is structured as follows. Section 2 introduces definitions and previous results. In Section 3 the main results are proved. In Conclusion we shortly analyze the necessary steps for development of the theory.

2 Mathematical model. Basic definitions and previous results

Let's consider the following mathematical model. Let X_i, $i = 1, 2, ..., n,...$, be a sequence of finite sets, $\prod_{i=1}^{n} X_i$ be a Cartesian product of X_i, $i = 1, 2, ..., n$, X^∞ be a set of all sequences where i-th element belongs to X_i. Define $\mathcal{A}$ be a σ-algebra on X^∞, generated by cylindrical sets. $\mathcal{A}$ is also Borel σ-algebra in Tychonoff product X^∞, where X_i, $i = 1, 2, ..., n,...$, has a discrete topology Bourbaki[2], Prokhorov and Rozanov[7]. On $(X^\infty, \mathcal{A})$ a probability measure P is defined. For any $n = 1, 2, ...$, assume that probability distribution P_n is a projection of measure P on the first n coordinates of random sequences from X^∞. It is clear that for every $B_n \subseteq \prod_{i=1}^{n} X_i$

$$P_n(B_n) = P(B_n \times X_n^\infty), \tag{1}$$

where $X_n^\infty = \prod_{i=n+1}^{\infty} X_i$.

Let $D_n(P)$ be the support of a measure P_n (Prokhorov and Rozanov[7]):

$$D_n(P) = \{\overline{x}_n \in \prod_{i=1}^{n} X_i,\ P_n(\overline{x}_n) > 0\}.$$

Define cylindrical set $\Delta_n(P)$ as follows:

$$\Delta_n(P) = D_n(P) \times X_n^{\infty}.$$

The sequence of cylindrical sets $\Delta_n(P)$, n=1,2,..., is not increasing and

$$\Delta(P) = \lim_{n\to\infty} \Delta_n(P) = \bigcap_{n=1}^{\infty} \Delta_n(P). \tag{2}$$

The set $\Delta(P)$ is a compact and is a support of measure P.

Let $\overline{x}_k \in \prod_{i=1}^{k} X_i$ and $\widetilde{x}_{k-1}$ is obtained from $\overline{x}_k$ by dropping the last coordinate.

Definition 1. Ban of measure P_n (Grusho *et al.*[3]) is a vector $\overline{x}_k \in \prod_{i=1}^{k} X_i$, $k \le n$, such that

$$P_n(\overline{x}_k \times \prod_{i=1}^{n-k} X_i) = 0.$$

Definition 2. Ban $\overline{x}_k$ of measure P_n is the smallest ban of measure P (Grusho *et al.*[3]) if

$$P_{k-1}(\widetilde{x}_{k-1}) > 0.$$

If $\overline{x}_k$ is a ban of measure P_n then for every $k \le s \le n$ and for every sequence $\overline{x}_k$ starting with $\overline{x}_s$ we have

$$P_s(\overline{x}_s) = 0.$$

If there exists a vector $\overline{x}_n \in \prod_{i=1}^{n} X_i$ such that $P_n(\overline{x}_n) = 0$, then there exists the smallest ban which is defined by the values of the first coordinates of vector $\overline{x}_n$.

Further under Λ_n we will understand a set of the smallest bans of measure P, which have lengths equal to n.

We also consider a set of probability measures Q_θ, $\theta \in \Theta$, on $(X^\infty, \mathcal{A})$ for which $Q_{n,\theta}$, $D_n(Q_\theta)$, $\Delta_n(Q_\theta)$, $\Delta(Q_\theta)$ are defined.

Consider the sequence of criteria for testing of hypotheses $H_{0,n} : P_n$ against alternatives $H_{1,n} : \{Q_{n,\theta},\ \theta \in \Theta\}$, $n = 1, 2, ...$

The statistical criterion for testing $H_{0,n}$ against $H_{1,n}$ is defined by a critical set S_n of vectors of length n. When the observed vector is in S_n then it leads to the acceptance of alternative $H_{1,n}$. If all vectors from S_n are bans of measure P_n, say that the criterion is defined by bans of measure P. Note that for every n we have $P_n(S_n) = 0$, if S_n is defined by bans.

Let $W_n(\theta)$ be the power function of criterion for testing $H_{0,n}$ against $H_{1,n}$. It is known that $W_n(\theta) = Q_{n,\theta}(S_n)$, $\theta \in \Theta$.

The basic problem considered in the paper is to find conditions when there exists such N that for all $n > N$ the power function $W_n(\theta) = 1$ for all $\theta \in \Theta$.

3 Mathematical results

Let's consider a set of probability measures $\{Q_\theta,\, \theta \in \Theta\}$ on $(X^\infty,\, \mathcal{A})$, for which $Q_{\theta,\,n}$, $D_n(Q_\theta)$, $\Delta_n(Q_\theta)$, $\Delta(Q_\theta)$ are defined as in Section 2.

Consider a problem of testing a sequence of hypotheses $H_{0,\,n} : P_n$ against complex alternatives $H_{1,\,n} : \{Q_{\theta,\,n},\, \theta \in \Theta\}$. Let $W_n(\theta)$ be a power function.

Theorem 1. *There exists a sequence of criteria for testing $H_{0,\,n}$ against $H_{1,\,n}$ with critical sets S_n, $n = 1,\, 2...$, defined by bans, for which exists such N that for every $n \geq N$ the power function $W_n(\theta) = 1$ if and only if there exists a closed set Δ such that for every θ, $\theta \in \Theta$,*

$$\Delta(Q_\theta) \subseteq \Delta,$$

and

$$\Delta(P) \cap \Delta = \emptyset.$$

Proof. The proof of the theorem 1 is carried out similar to the proving of the theorems 1 and 2 in Grusho *et al.*[6]. We will give the main ideas of the proof which will be required in further reasonings.

Let's define

$$\sigma = \bigcup_{k=1}^{\infty} \bigcup_{\overline{x}_k \in \Lambda_k} I(\overline{x}_k),$$

where $I(\overline{x}_k)$ be the elementary cylindrical set in X^∞, which is generated by the smallest ban $\overline{x}_k$.

It is easy to prove that

$$\sigma = X^\infty \backslash \Delta(P).$$

By the condition of the theorem 1 $\Delta \subseteq \sigma$. In the considered case the topological Tychonoff product X^∞ is a compact space Bourbaki[2]. Then for the set Δ there exists a finite cover

$$\sigma_N = \bigcup_{k=1}^{N} \bigcup_{\overline{x}_k \in \Lambda_k} I(\overline{x}_k).$$

The set σ_N is a cylindrical set and $\sigma_N = C_N \times X_N^\infty$. From the conditions of the theorem 1 and formula (1) it follows that for every $\theta \in \Theta$

$$1 = Q_\theta(\Delta(Q_\theta)) \leq Q_\theta(\sigma_N) = Q_{\theta,\,N}(C_N).$$

If critical set S_N of criterion for testing $H_{0,\,N} : P_N$ against the complex alternatives $H_{1,\,N} : \{Q_{\theta,\,N},\, \theta \in \Theta\}$ is chosen as $S_N = C_N$, then $W_n(\theta) = 1$ for all θ. The sufficiency is proved.

Let' prove the necessity. Let S_N be such a critical set for which $Q_{\theta,\,N}(S_N) = 1$ for all θ, $\theta \in \Theta$, and S_N is defined by bans of measure P. Then for all θ, $\theta \in \Theta$, the set $D_N(Q_\theta) \subseteq S_N$, and using the definition of σ_N we conclude that every set S_N defined by bans satisfies to

$$S_N \times X_N^\infty \subseteq \sigma_N.$$

For cylindrical sets we have

$$\Delta_N(Q_\theta) \subseteq \sigma_N,\ \theta \in \Theta.$$

Thus

$$\Delta(Q_\theta) \subseteq \sigma_N,\ \theta \in \Theta.$$

By the definition

$$S_N \cap D_N(P) = \emptyset.$$

As

$$\Delta(P) \subseteq \Delta_N(P).$$

then it follows that

$$\sigma_N \cap \Delta(P) = \emptyset.$$

σ_N is a cylindrical set and in discrete topology on $\prod_{i=1}^{N} X_i$ it is a closed set. So we can define $\Delta = \sigma_N$. The theorem 1 is proved.

Example 1. Let $x \in X^\infty$. For every n denote $x|_n$ be a vector of the length n which coincides with value of the first n coordinates of the sequence x. Assume that $P(x) = 1$.

Define two alternatives. Let y be a sequence of X^∞ such that $x|_n = y|_n$, but $x|_{n+1} \neq y|_{n+1}$. Define $Q_1(y) = 1$. Let z be a sequence from X^∞, $z|_m = x|_m$ and $z|_{m+1} \neq x|_{m+1}$ and $m > n$. Define $Q_2(z) = 1$.

It is clear that the set of bans for P_k contains all vectors from $\prod_{i=1}^{k} X_i$ which are different from $x|_k$. For $k \leq n$ the support of $Q_{1,\,k}$ is not covered by bans of measure P. But when $k = n + 1$ bans of measure P cover the support of $Q_{1,\,n+1}$. The analogical conclusions can be derived for measure Q_2.

That is why for $k \geq m + 1$ the bans of measure P cover supports of all projections of measures Q_1 and Q_2. It is clear that number of steps which are needed to solve the problem of testing with the help of tests defined by bans and power function equals to 1 is $m + 1$.

Example 2. Let's assume that with the help of a channel rows of a relational data base are transmitted. Let the domains of attributes $A_1, ..., A_r$ equal to $X_1, ..., X_r$. This is the case when the model is described by the sequence of different alphabets. An example of a ban here can be described by the changing of values of attributes. For example, a change of an attribute A_1 in neighbor rows should be supported by the change of the value of attribute A_2. Otherwise we get a ban in the sequence.

Generalizing the example 1 we can define an algorithm of constructing the test with power function equals to 1.

Assume that the sufficient conditions of the theorem 1 are fulfilled. That's why we know the closed set Δ. In the considered topological space every closed set can be represented as follows

$$\Delta = \bigcap_{n=1}^{\infty} I_n,$$

where $I_n = C_n \times X_n^\infty$, $n = 1, 2, ...$, is not increasing sequence of cylindrical sets. According to the theorem 1 the needed test will be constructed when bans of

measure P cover the set C_n for some n. Due to finiteness of the model for every n we can use the brute force algorithm by comparing bans of measure P_n with elements of C_n. Due to the result of the theorem 1 we should consider only finite number of n to construct the critical set with power function equals to 1. Note that for every $n: P_n(C_n) = 0$.

Of course there may be a lot of cases when the complexity of the algorithm can be reduced in comparison with brute force algorithm. But these cases should be considered autonomously.

In all cases the construction of the test is the preliminary work and its usage can be described by an efficient algorithm.

4 Conclusion

We defined the requirements for a monitoring system of some technological process which separates normal and abnormal behavior of the process. It's necessary to prevent false alarms and for sure to find the process deviations from the normal behavior. Usage of bans helps to exclude false alarms by definition.

Theorem 1 defines the conditions when there exist tests defined by bans with power functions equal to 1. If conditions are fulfilled we define brute force algorithms for the construction of such tests defined by bans which can be implemented in a monitoring system.

The most difficult work should be done at the stage of preparation of monitoring system for starting. There are a lot of problems concerning the reduction of complexity of preliminary stage and with understanding of behavior of monitoring system when preliminary data are incomplete.

4.1 Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grant 13-01-00215).

References

1. S. Axelson. The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection, *in: Proc. of the 6th Conference on Computer and Communications Security*, 1999.
2. N. Bourbaki *Topologie G'en'erale. Russian translation*, Science, Moscow, 1968.
3. A. Grusho, N. Grusho, and E. Timonina. Consistent sequences of tests defined by bans, *Springer Proceedings in Mathematics and Statistics, Optimization Theory, Decision Making, and Operation Research Applications*, Springer, 281–291, 2013.
4. A. Grusho and E. Timonina. Prohibitions in discrete probabilistic statistical problems, *Discrete Mathematics and Applications*, 21, **3**, 275–281, 2011.
5. A. Grusho, N. Grusho, and E. Timonina. Generation of Probability Measures with the Given specification of the Smallest Bans, in: *Proceedings of 28^{th} European Conference on Modelling and Simulation*, (May 27-30, 2014, Brescia, Italy), Digitaldruck Pirrot GmbHP, Dudweiler, Germany, 565–569, 2014.

6. A. Grusho, N. Grusho, and E. Timonina. Power functions of statistical criteria defined by bans, in: *Proceedings of 29th European Conference on Modelling and Simulation*, (May 26-30, 2015, Varna, Bulgaria), Digitaldruck Pirrot GmbHP (to appear), 2015.
7. U.V. Prokhorov and U.A. Rozanov. *Theory of probabilities*, Science, Moscow, 1993.

On some issues in trajectory modeling with finite mixture models

Jean-Daniel Guigou[1], Bruno Lovat[2], and Jang Schiltz[3]

[1] University of Luxembourg, LSF, 4, rue Albert Borschette, L-1246 Luxembourg, Luxembourg
(E-mail: jean-daniel.guigou@uni.lu)
[2] University of Lorraine, BETA, 13 place Carnot, F- 54035 Nancy, France
(E-mail: bruno.lovat@univ-lorraine.fr)
[3] University of Luxembourg, LSF, 4, rue Albert Borschette, L-1246 Luxembourg, Luxembourg
(E-mail: jang.schiltz@uni.lu)

Abstract. We explore some issues arising in trajectory modeling with finite mixture models adding covariates to the trajectory specifications. More specifically, we use a generalized model that allows allows non parallel trajectories for different values of covariates and discuss the use of covariant variables which have no significant influence on the group membership probabilities, as well as the risk of multicolinearity in the case of time-varying covariates. We illustrate our discussion by giving typical salary curves for the employees in the private sector in Luxembourg between 1987 and 2006, as a function of their country of residence, as well as of Luxembourg's consumer price index (CIP).
Keywords: Trajectory Modeling, Generalized Finite Mixture Models, Data Illustration.

1 Introduction

Time series are largely used in economics, sociology, psychology, criminology and medicine and a host of statistical techniques are available for analyzing them (see Singer and Willet[18]). In that context, the study of developmental trajectories is a central theme. In the 1990s, the generalized mixed model assuming a normal distribution of unobserved heterogeneity (Bryk and Raudenbush[1]), multilevel modeling (Goldstein[5]), latent growth curves modeling (Muthén[11], Willett and Sayer[22]) and the nonparametric mixture model, based on a discrete distribution of heterogeneity (Jones, Nagin & Roeder[10]) have emerged. There has been a growing interest in this approach to answer questions about atypical subpopulations (see Eggleston, Laub and Sampson[4]).

Latent class growth analysis, also called nonparametric mixed model or semiparametric mixture model was originally discussed by Nagin and Land[14], Nagin[12] and Roeder, Lynch and Nagin[16] and is specifically designed to detect the presence of distinct subgroups among a set of trajectories and repre-

16^{th} *ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

sents an interesting compromise between analysis around a single mean trajectory and case studies (von Eye & Bergman[20]). Compared to subjective classification methods, the nonparametric mixed model has the advantage of providing a formal framework for testing the existence of distinct groups of trajectories. This method does not assume a priori that there is necessarily more than one group in the population. Rather, an adjustment index is used to determine the number of sub-optimal groups. This is a significant advance over other categorical methods which determine the number of groups only subjectively (von Eye & Bergman[20]).

While the conceptual aim of the analysis is to identify clusters of individuals with similar trajectories, the model's estimated parameters are not the result of a cluster analysis but of maximum likelihood estimation (Nagin[13]). Moreover, this method allows to evaluate the accuracy of the assignement of the individuals to the different sub-groups and to consider the variation of this accuracy in subsequent analyses (Dupéré et al.[3]). Nagin and Odgers[15] document numerous applications of group-based trajectory modeling in criminology and clinical research. They state that the appeal of group-based trajectory modeling for the future lies in the potential for the innovative application of trajectory models on their own, in conjunction with other statistical methods or embedded within creative study designs while carefully considering the perils and pitfalls inherent in the use of any methodology.

Nagin[13] develops a generalization of his model that allows trajectories depending on covariates. Unfortunately in this model, trajectories for different values of the covariates are constrained to remain always parallel and the error terms of the different groups all have the same dispersion. Schiltz[17] finally presents a generalization of the model that overcomes these two weaknesses. In this paper we discuss some remaining issues and illustrate them by data examples.

It is structured as follows. In section two, we present the basic version of Nagin's finite mixture model, as well as two of his generalizations. In section three, we present the generalization of Schiltz and some of its statistical properties. In section four, finally, we highlight some remaining issues by means of a data example from economics.

2 Nagin's Finite Mixture Model

Starting from a collection of individual trajectories, the aim of Nagin's finite mixture model is to divide the population into a number of homogenous subpopulations and to estimate, at the same time, a typical trajectory for each sub-population (Nagin[13]).

More, precisely, consider a population of size N and a variable of interest Y. Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable Y, taken at times $t_1, ..., t_T$ for subject number i. To estimate the parameters defining the shape of the trajectories, we need to fix the number r of desired subgroups. Denote the probability of a given subject to belong to group number j by π_j.

The objective is to estimate a set of parameters $\Omega = \{\pi_j, \beta_0^j, \beta_1^j, ...; j = 1, ..., r\}$ which allow to maximize the probability of the measured data. The

particular form of Ω is distribution specific, but the β parameters always perform the basic function of defining the shapes of the trajectories. In Nagin's finite mixture model, the shapes of the trajectories are described by a polynomial function of age or time. In this paper, we suppose that the data follow a normal distribution. Assume that for a subject in group j

$$y_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k + \varepsilon_{it}, \tag{1}$$

where s denotes the order of the polynomial describing the trajectories in group j and ε_{it} is a disturbance assumed to be normally distributed with a zero mean and a constant standard deviation σ. If we denote the density of the standard centered normal law by ϕ and $\beta^j t_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k$, the likelihood of the data is given by

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi \left(\frac{y_{it} - \beta^j t_{it}}{\sigma} \right). \tag{2}$$

The disadvantage of the basic model is that the trajectories are static and do not evolve in time. Thus, Nagin introduced several generalizations of his model in his book (Nagin[13]). Among others, he introduced a model allowing to add covariates to the trajectories. Let $z_1, ..., z_M$ be M covariates potentially influencing Y. We are then looking for trajectories

$$y_{it} = \sum_{k=0}^{s} \beta_k^j t_{it}^k + \alpha_1^j z_1 + ... + \alpha_M^j z_M + \varepsilon_{it}, \tag{3}$$

where ε_{it} is normally distributed with zero mean and a constant standard deviation σ. The covariates z_m may depend or not upon time t.

A second model extension also allows the statistical testing of hypotheses about whether individual-level characteristics distinguish trajectory group membership (Nagin[13]). To that effect, group membership probability π_j is written as a function of a possible effect θ_j of variable x to this probability.

$$\pi_j(x_i) = \frac{e^{x_i \theta_j}}{\sum_{k=1}^{r} e^{x_i \theta_k}} \tag{4}$$

The actual statistical testing can then be done by conventional z-score based testing (Greene[7]) or the Wald test (Wald[21]).

But Nagin's generalized models still have two major drawbacks. First, the influence of the covariates in these model is unfortunately limited to the intercept of the trajectory. This implies that for different values of the covariates, the corresponding trajectories will always remain parallel by design, which does not necessarily correspond to reality.

Secondly, in Nagin's models, the standard deviation of the disturbance is the same for all the groups. That too is quite restrictive. One can easily

imagine situations in which in some of the groups all individual are quite close to the mean trajectory of their group, whereas in other groups there is a much larger dispersion.

3 A more flexible model

3.1 Definition

To address and overcome these two drawbacks, Schiltz[17] proposed the following generalization of Nagin's model.

Let $x_1 ... x_M$ and $z_{i_1}, ..., z_{i_T}$ be covariates potentially influencing Y. Here the x variables are covariates not depending on time like gender or cohort membership in a multicohort longitudinal study and the z variable is a covariate depending on time like being employed or unemployed. They can of course also designate time-dependent covariates not depending on the subjects of the data set which still influence the group trajectories, like GDP of a country in case of an analysis of salary trajectories.

The trajectories in group j will then be written as

$$y_{it} = \sum_{k=0}^{s} \left(\beta_k^j + \sum_{m=1}^{M} \alpha_{km}^j x_m + \gamma_k^j z_{i_t} \right) t_{it}^k + \varepsilon_{it}, \tag{5}$$

where the disturbance ε_{it} is normally distributed with mean zero and a standard deviation σ_j constant inside group j but different from one group to another. Since, for each group, this model is just a classical fixed effects model for panel data regression (see Woolridge[23]), it is well defined and we can get consistent estimates for the model parameters.

This model allows obviously to overcome the drawbacks of Nagin's models. The standard deviation of the uncertainty can vary across groups and the trajectories depend in a nonlinear way on the covariates. In practice this dependance of all the power coefficients of the polynomials may considerably extend the computation time for the parameters, so it can be useful just to work with a first or second order dependance instead of using the full model.

3.2 Statistical Properties

The model's estimated parameters are the result of maximum likelihood estimation. As such, they are consistent and asymptotically normally distributed (Cramèr[2]; Greene[7]; Theil[19]).

In our model, for a given group, the trajectories follow in fact a nonlinear regression model. As such, exact confidence interval procedures or exact hypothesis tests for the parameters are generally not available (Graybill and Iyer[6]). There exist however approximative solutions. The standard error can be approximated for instance by a first-order Taylor series expansion (Greene[7]).

This approximate standard error (ASE) is usually quite precise if the sample size is sufficiently large.

Consider model (5), for which $(2 + M)s$ regression parameters have to be estimated. Then confidence intervals of level α for the parameters β_k^j are just

$$CI_\alpha(\beta_k^j) = \left[\hat{\beta}_k^j - t_{1-\alpha/2;N-(2+M)s} ASE(\hat{\beta}_k^j); \hat{\beta}_k^j + t_{1-\alpha/2;N-(2+M)s} ASE(\hat{\beta}_k^j)\right], \tag{6}$$

where $t_{1-\alpha;n}$ denotes as usually the $1-\alpha$ quantile of the Student distribution with n degrees of freedom.

The confidence intervals for the α_{kl}^j and γ_k^j are obtained in the same way.

The confidence intervals of level α for the disturbance factor σ_j is given by

$$CI_\alpha(\sigma_j) = \left[\sqrt{\frac{(N-(2+M)s-1)\hat{\sigma_j}^2}{\chi^2_{1-\alpha/2;N-(2+M)s-1}}}; \sqrt{\frac{(N-(2+M)s-1)\hat{\sigma_j}^2}{\chi^2_{\alpha/2;N-(2+M)s-1}}}\right], \tag{7}$$

where $\chi^2_{1-\alpha;n}$ denotes the $1-\alpha$ quantile of the Chi-Square distribution with n degrees of freedom.

4 Empirical illustration

For the following data example, we use Luxembourg administrative data originating from the General Inspectorate of Social Security, IGSS (Inspection générale de la sécurité sociale). The data have previously been described and exploited with Nagin's basic model by Guigou, Lovat and Schiltz ([8], [9]). The file contains the salaries of all employees of the Luxembourg private sector who started their work in Luxembourg between 1980 and 1990 at an age of less than 30 years. This choice was made to eliminate people with a long carrier in another country before moving to Luxembourg. The main variables are the net annual taxable salary, measured in constant (2006 equivalent) euros, gender, age at first employment, residentship and nationality, sector of activity, marital status and the years of birth of the children. The file consists of 1303010 salary lines corresponding to 85049 employees. In Luxembourg, the maximum contribution ceiling on pension insurance is 5 times the minimum wage, currently 7577 EUR (2006 equivalent euros) per month. Wages in our data are thus also capped at that number.

We will not present here an exhaustive analysis of the whole dataset, but just two illustrations of the possibilities of our generalized mixture model and its differences from Nagin's model. We concentrate on the first 20 years of the careers of the employees who started working in Luxembourg in 1987. That gives us a sample of 1716 employees. We will first compute typical salary trajectories for them, taking into account the country of residence of the employees and then typical salary trajectories as a function of the Luxembourg Consumer Price Index (CPI).

4.1 First data example

Here we illustrate the case of a covariant variable which has no significant influence on the group membership probabilities and hence on the constitution of the groups. We analyze if the fact to be either a Luxembourg resident or a commuter has an influence on the salary. Unlike the case of the gender (Schiltz[17]), this covariate does not distinguish trajectory membership. That means that Nagin's model does not provide different trajectories for residents and non residents. Hence Nagin's model implies that throughout the 20 first years of their career residents and non residents have the same salary trajectories. To constrain it to give us different solutions for residents and commuters, we took the 6 group solution resulting from Nagin's generalized model and computed the salary trajectories in each group separately for residents and non residents. The result is shown in Fig. 1. It shows the salary of employees

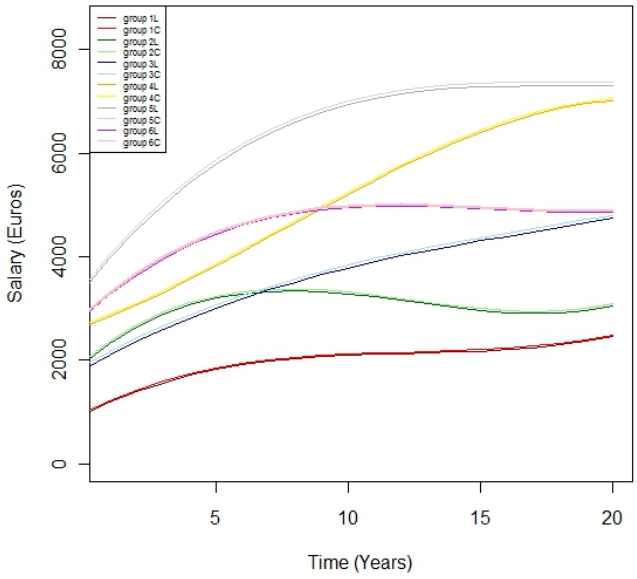

Fig. 1. Salary evolution by country of residence, modelized by Nagin's model

in Luxembourg during the first 20 years of their professional career. The curves in dark colors (groups 1L, 2L, 3L, 4L, 5L and 6L) represent employees living in Luxembourg, the curves in light colors represent commuters living in the surrounding countries. Since the country of residence has no significant effect on the salary, it is not astonishing that in each group the curves for residents and commuters are very close. But due to the limitations of the model, the evolution of the salaries seems to be exactly the same for both categories. In

fact the commuters seem to earn a little more money then the residents in all six groups throughout the whole time line.

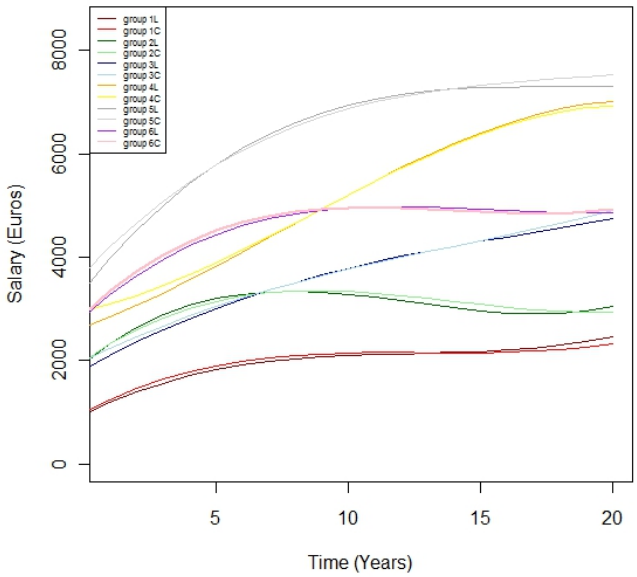

Fig. 2. Salary evolution by country of residence, modelized by our model

Fig. 2 shows the six group solution for the 20 first year of Luxembourg employees calibrated with our model. Here we see quite a different pattern. The difference between residents and commuters is of course not very large, but the evolution is not parallel any more. In the two low salary groups Luxembourg residents earn a little bit more during the last few years, whereas commuters have a slightly higher salary during the first years of their career. In the two middle salary groups Luxembourg residents earn a little bit less at the beginning and the end, but slightly more in the middle of the trajectory. The same situation can be observed for the group with the highest salaries, whereas for the second highest salary group, its the Luxembourg residents that gain more at the beginning and the end, but less in the middle of our time interval. Hence we see, that our model allows a more precise description of reality than Nagin's ones.

We obtained this results by calibrating the model

$$S_{it} = (\beta_0^j + \alpha_0^j x_i) + (\beta_1^j + \alpha_1^j x_i)t + (\beta_2^j + \alpha_2^j x_i)t^2 + (\beta_3^j + \alpha_3^j x_i)t^3, \quad (8)$$

where S denotes the salary and x the country of residence variable (The Luxembourg resident are coded by 1 and the commuters by 0). Table 1 shows the values of the parameters for a 6-group solution.

Results for group 1

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	950.373	41.733	861.903	1038.843
α_0	34.576	3.402	29.484	39.669
β_1	259.220	16.792	223.623	294.817
α_1	17.415	0.966	15.366	19.464
β_2	-19.812	1.834	-23.701	-15.923
α_2	-1.805	0.106	-2.028	-1.581
β_3	0.531	0.058	0.410	0.653
α_3	0.026	0.003	0.018	0.033

Results for group 2

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	1946.965	47.927	1845.365	2048.565
α_0	33.799	11.426	9.576	58.022
β_1	420.544	19.284	379.664	461.424
α_1	-55.339	4.598	-65.086	-45.593
β_2	-39.272	2.107	-43.739	-34.806
α_2	9.235	0.502	8.170	10.300
β_3	1.050	0.066	0.910	1.190
α_3	-0.342	0.016	-0.375	-0.309

Results for group 3

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	1833.225	71.228	1682.228	1984.221
α_0	150.383	12.588	123.761	177.006
β_1	284.880	28.659	224.124	345.635
α_1	-26.094	5.053	-36.806	-15.382
β_2	-11.063	3.131	-17.700	-4.426
α_2	0.520	0.552	-0.650	1.690
β_3	0.204	0.098	-0.004	0.412
α_3	0.040	0.018	0.003	0.077

Results for group 4

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	2645.589	73.032	2490.767	2800.411
α_0	307.435	21.509	249.121	365.750
β_1	188.837	29.386	126.542	251.132
α_1	-69.719	11.068	-93.183	-46.256
β_2	12.030	3.210	5.224	18.836
α_2	5.112	1.209	2.548	7.675
β_3	-0.528	0.101	-0.741	-0.314
α_3	-0.130	0.038	-0.211	-0.050

Results for group 5

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	3361.231	134.999	3075.047	3647.415
α_0	310.893	17.475	347.938	273.849
β_1	649.017	54.318	533.867	764.167
α_1	-101.596	7.031	-86.691	-116.501
β_2	-35.623	5.934	-48.203	-23.043
α_2	7.832	0.768	9.460	6.203
β_3	0.651	0.186	0.256	1.045
α_3	-0.148	0.024	-0.097	-0.199

Results for group 6

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	2853.326	64.784	2715.991	2990.662
α_0	38.171	28.242	-21.699	98.040
β_1	457.569	26.067	402.310	512.828
α_1	27.307	11.365	3.217	51.396
β_2	-31.612	2.848	-37.649	-25.575
α_2	-4.920	1.242	-7.552	-2.289
β_3	0.688	0.089	0.499	0.877
α_3	0.180	0.039	0.098	0.263

Table 1. Parameter estimates for model 8

We observe that all parameters are significant, with the exception of α_2 and β_3 for group 3 and α_0 for group 6. Hence, there really seems to be a nonlinear relationship between the salaries and the country of residence and a simple parallel shift is not enough to explain what is going on.

The disturbance terms for the six groups are $\sigma_1 = 39.31$, $\sigma_2 = 49.11$, $\sigma_3 = 70.96$, $\sigma_4 = 79.45$, $\sigma_5 = 115.6$ and $\sigma_6 = 46.38$ respectively. The dispersion varies from group to group and is generally higher in the groups with higher salaries then in those with lower salaries.

4.2 Second data example

In Luxembourg, the establishment of consumer price indices (CPI) started at the beginning of the 1920s. The current index, which came into effect in January 1997, complies with the Community regulations concerning the harmonised consumer price index except that its weighting excludes the consumer spending of non-residents on Luxembourg territory.

The index is issued monthly. It is presented in the form of a chain index, the weighting of which is adjusted every year to take account of modifications in households' consumer habits. Calculated for the twelve months of a given year compared to the month of December of the previous year, it is published on base 100 at 1st January 1948 for the purposes of the sliding wage scale.

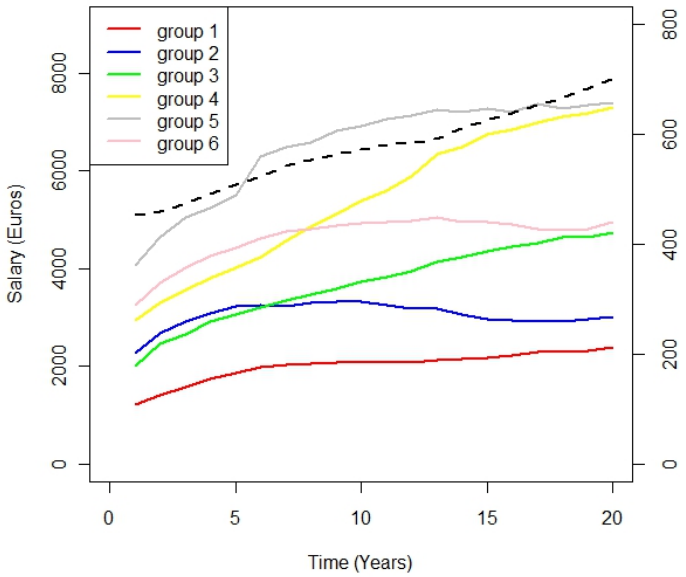

Fig. 3. Typical salary trajectories for the 6 group solution and evolution of Luxembourg's CPI (dashed line).

Fig. 3 shows the salary trajectories of the 6 groups (scale at the left side of the y axis), as well as the CPI of Luxembourg (in dashed black, scale on the right side of the y axis) during the same time. The two first groups contain the employees that gain the legal minimum wage and the minimal qualified wage respectively. Groups three and four represent middle class salaries and groups five and six the higher salary groups. Moreover, groups two and six represent employees with rather flat careers. Their salary is more or less constant from year five on. They are just distinguished by their starting salary. Groups three, four and five on the other hand represent more dynamical careers, again characterized mainly by the differences in their starting salaries.

Our example illustrates the dependence of the trajectories on time varying covariates. We use the same data as before and analyze the influence of Luxembourg's CPI on the salary trajectories. Here we have to be careful in our model choice because the two time series have a correlation of 0.995. Hence a model like

$$S_{it} = (\beta_0^j + \gamma_0^j z_{it}) + (\beta_1^j + \gamma_1^j z_{it})t + (\beta_2^j + \gamma_2^j z_{it})t^2 + (\beta_3^j + \gamma_3^j z_{it})t^3, \qquad (9)$$

where S denotes the salary and z_t is Luxembourg's CPI in year t of the study, makes no sense. Because of obvious multicolinearity problems, almost none of the parameters would be significant.

Therefore, we simplify the model and calibrate

$$S_{it} = (\beta_0^j + \gamma_0^j z_{it}) + \gamma_1^j z_{it}t + \gamma_2^j z_{it}t^2 + \gamma_3^j z_{it}t^3. \qquad (10)$$

Tables 2 shows the values of the parameters for a 6 group solution.

We observe a significant influence of the CPI for all six groups, which is not astonishing, since by law, the salaries are coupled with the CPI. For groups two, three and six all parameters are significant. The trajectories in groups one and five do not have any constant term, nor a linear dependency on the CPI but depend only on the interaction of CPI and time. Group four, finally, exhibits only linear behaviour with respect to CPI, as well as the interaction of CPI and time.

The disturbance terms for the six groups are $\sigma_1 = 41.49$, $\sigma_2 = 33.18$, $\sigma_3 = 68.48$, $\sigma_4 = 64.84$, $\sigma_5 = 111.83$ and $\sigma_6 = 39.74$ respectively. Again, we observe that the minimal wage group exhibits the smallest variability, whereas the high salary groups four and five also have the highest disturbance term.

5 Conclusion

In this article, we presented Nagin's finite mixture model and Schiltz's generalization a key characteristic of which is its ability to modelize nearly all kind of trajectories and to add covariates to the trajectories themselves in a nonlinear way.

We used this model to illustrate some issues about trajectory modeling through a data example related to salary trajectories. In the first part, we

Results for group 1

Parameter	Estimate	Standard error	95% confidence intervals	
			Lower	Upper
β_0	321.381	1189.430	-2213.502	2856.093
γ_0	1689.492	277.834	-4.232	7.611
γ_1	0.400	0.120	0.143	0.656
γ_2	-0.034	0.007	-0.049	-0.019
γ_3	0.0008	0.0002	0.0005	0.0013

Results for group 2

Parameter	Estimate	Standard error	95% confidence intervals	
			Lower	Upper
β_0	7688.158	951.103	5660.197	9714.832
γ_0	-13.095	2.222	-17.822	-8.350
γ_1	1.260	0.096	1.055	1.465
γ_2	-0.097	0.006	-0.109	-0.085
γ_3	0.0025	0.0002	0.0022	0.0028

Results for group 3

Parameter	Estimate	Standard error	95% confidence intervals	
			Lower	Upper
β_0	682.638	196.327	2641.924	1101.045
γ_0	-11.367	4.586	-21.135	-1.586
γ_1	0.983	0.199	0.559	1.406
γ_2	-0.048	0.012	-0.073	-0.023
γ_3	0.0010	0.0003	0.0003	0.0017

Results for group 4

Parameter	Estimate	Standard error	95% confidence intervals	
			Lower	Upper
β_0	8473.081	1859.349	4511.016	12434.892
γ_0	-13.083	4.342	-22.335	-3.825
γ_1	0.927	0.188	0.527	1.328
γ_2	-0.013	0.011	-0.036	0.010
γ_3	-0.0003	0.0003	-0.0009	0.0004

Results for group 5

Parameter	Estimate	Standard error	95% confidence intervals	
			Lower	Upper
β_0	4798.276	3205.141	-2034.302	11630.238
γ_0	-2.846	7.488	-18.806	13.115
γ_1	1.315	0.324	0.0624	2.006
γ_2	-0.081	0.019	-0.122	-0.040
γ_3	0.0016	0.0005	0.0005	0.0027

Results for group 6

Parameter	Estimate	Standard error	95% confidence intervals	
			Lower	Upper
β_0	8332.439	1139.127	5903.348	10759.713
γ_0	-12.472	2.661	-18.145	-6.800
γ_1	1.378	0.015	1.132	1.623
γ_2	-0.094	0.007	-0.108	-0.079
γ_3	0.0022	0.0002	0.0018	0.0026

Table 2. Parameter estimates for model 10

showed that the generalized model from Schiltz[17] is able to detect different behaviour for different values of the covariant even in the case of covariant variables which have no significant influence on the constitution of the clusters. In the second part, we showed that models with time dependent covariates have to be considered carefully, since colinearity problems between the covariate and time can lead to parameters that are no more interpretable. In that case, it is necessary to take out the time dependency of the trajectories in order to preserve the dependency on the covariate.

References

1. A.S. Bryk and S.W. Raudenbush. *Hierarchical linear models*, Newbury Park, CA: Sage, 1992.
2. H. Cramér. *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press, 1946.
3. V. Dupéré, E. Lacourse, F. Vitaro and R.E. Tremblay. Méthodes d'analyse du changement fondées sur les trajectoires de développement individuel: modèles de régression mixtes paraméetriques et non paramétriques. *Bulletin de mthodologie sociologique*, 95, 26–57, 2007.
4. E.P. Eggleston, J.H. Laub and R.J. Sampson. On the Robustness and Validity of Groups. *Journal of Quantitative Criminology*, 20(1), 37–42, 2004.
5. H. Goldstein. *Multilevel Statistical Models*. London: Arnold, 1995.
6. F.A. Graybill and H.K. Iyer. *Regression Analysis: Concepts and Applications*. Belmont,CA: Duxbury Press, 1994.
7. W.H. Greene. *Econometric Analysis*. Seventh Edition, Harlow: Pearson Education, 2012.
8. J.-D. Guigou, B. Lovat and J. Schiltz. The impact of ageing population on pay-as-you-go pension systems: The case of Luxembourg. *Journal of International Finance and Economics*, 10(1), 110–122, 2010.
9. J.-D. Guigou, B. Lovat and J. Schiltz. Optimal mix of funded and unfunded pension systems: the case of Luxembourg. *Pensions*, 17(4), 208–222, 2012.
10. B.L. Jones, D.S. Nagin and K. Roeder. A SAS Procedure Based on mixture Models for Estimating Developmental Trajectories. *Sociological Methods & Research*, 29(3), 374–393, 2001.
11. B.O. Muthén. Latent Variable Modeling in Heterogeneous Populations. *Psychometrika*, 54(4), 557–585, 1989.
12. D.S. Nagin. Analyzing Developmental Trajectories: Semi-parametric. Groupe-based Approach. *Psychological Method*, 4, 139–157, 1999.
13. D.S. Nagin. *Group-Based Modeling of Development*. Cambridge, MA: Harvard University Press, 2005.
14. D.S. Nagin and K.C. Land. Age, criminal careers and population heterogeneity: Specifiction and estimation of a nonparametric, mixed Poisson model. *Criminology*, 31, 327–362, 1993.
15. D.S. Nagin and C.L. Odgers. Group-Based Trajectory Modeling (Nearly) Two Decades Later. *Journal of Quantitative Criminology*, 26, 445–453, 2010.
16. K. Roeder, K.G. Lynch and D.S. Nagin. Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association*, 94, 766–776, 1999.
17. J. Schiltz. A generalization of Nagin's finite mixture model. In: M. Stemmler, A. Eye and W. Wiedermann (Eds.) *Dependent data in social sciences research: Forms, issues, and methods of analysis*, 2015.

18. J.D. Singer, and J.B. Willet. *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press, 2003.
19. H. Theil. *Principles of Econometrics*. New York, NY: Wiley, 1971.
20. A. von Eye and L.R. Bergman. Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. *Development and Psychopathology*, 15, 553–580, 2003.
21. A. Wald. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observation Is Large. *Transactions of the American Mathematical Society*, 54, 426–482, 1943.
22. J.B. Willett and A.G. Sayer. Using Covariance Structure Analysis to Detect Correlates and Predictors of Individual Change Over Time. *Psychological Bulletin*, 116(2), 363–381, 1994.
23. J. Woolridge. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT press, 2002.