Compositional Data Analysis-Coherent

Forecasting Mortality Model With

Cohort Effect

Amos BATIONO1, Leo ODONGO2, Karim DERRA3

1 Pan African University, Institute of Basic Science Technology and Innovation, Nairobi, Kenya

2 Department of Mathematics and Actuarial Science, Kenyatta University, Nairobi, Kenya

3Clinical Research Unit of Nanoro, Institute of Research in Health Science, Burkina Faso

1 [batimosa92@gmail.com](mailto:batimosa92@gmail.com), 2 [Odongo.leo@ku.ac.ke](mailto:Odongo.leo@ku.ac.ke), 3 [kderra@crun.bf](mailto:kderra@crun.bf)

**ABSTRACT**

In this paper, an extension of the Coherent forecasts of mortality with compositional data analysis (CoDa) model of Bergeron-Boucher et al. (2017) to cohort effect is proposed applied to data from six African countries. The process of fitting this model starts by adapting the Renshaw and Haberman (2006) to compositional data analysis (CODA) as suggested by Bergeron-Boucher et al. (2017). The proposed CoDa-cohort model generally fits the data better than the original cohort model of Renshaw and Haberman (2006). In order to get the full CoDa-cohort-coherent model the multiple population factor is included in CoDa-cohort model. Then a comparison between CoDa -coherent and CoDa-cohort-coherent models revealed that the latter model generally gave a better fit to the data based on Aitchinson distance (AD).

Key words: Mortality, Compositional data analysis, coda, Coherent, Cohort, Forecast

1. **INTRODUTION**

Mortality forecasting has become almost unavoidable in many fields such as financial, actuarial and health care institutions. A number model exists in this field. But the most famous belongs to the group of extrapolative forecasting models, the Lee and Carter (1992) model which is widely used and many other models are based on it due to its accuracy and simplicity. However, its limitations have made it subject of further modifications and extensions in order to overcome some flaws or to adapt it to some specific situations or data.

Lee and Li (2005) proposed an extension to forecast mortality of many population simultaneously and coherently. Indeed, the original Lee Carter model was meant to forecast a single population mortality rate and so it could not take into account two populations male and female simultaneously, for instance, while forecasting. When the forecasts of populations do not diverge over the long run, it is said to be coherent. To avoid that divergence, Lee and Li (2005) used the notion of common factor that captures the effect induced by the common characteristics of all the populations on the rate of mortality.

Sometime mortality rate is affected by the year of birth. It can be a consequence of an event that occurred in that specific year. To capture this effect on a given cohort (year of birth) Renshaw and Haberman (2006) added an age–period–cohort term to the original Lee and Carter model.

In 2008 Oeppen adapted Lee and Cartel original model to compositional data. Compositional data can be seen as vectors of components those values are strictly positive with sum equals to a constant. The space of such kind of data is called simplex space. For that, standard statistical analysis is affected (Aitchison 1986, Bergeron-Boucher et al., 2017). Since life table death distribution is under the same constraint, Oeppen (2008) proposed to treat using CoDa framework. This framework provides tools especially for these kinds of data. Oeppen (2008) used the centred log-ratio transformation to bring the data out of the simplex space to the real space before starting analysing them. The advantage of this transformation is that it conserves distance from the simplex to Euclidean spaces.

Recently Bergeron-Boucher et al. (2017) proposed a coherent forecasting model for compositional data. The authors extended the Oeppen(2008) model to coherent forecasting by using the notion of common factor used in Lee and Li (2005). CoDa model solves some problems of Lee and Carter original model such as fixed rate of mortality improvement (RMI) (Bergeron-Boucher et al., 2017). Indeed, many studies have shown that the assumption of fixed RMI is not supported by empirical studies (Yue et al., 2008). The CoDa coherent model has some advantages but it does not take into account the cohort effect which when ignored can lead to misleading results.

This study is focused on including a cohort effect into Bergeron-Boucher et al. (2017) model. In the first part of this article, we adapt the model of Renshaw and Haberman (2006) to CoDa; in the second we will include the cohort effect in the CoDa-coherent model of Bergeron-Boucher et al. (2017); finally, the models will be applied to real mortality data of some African countries.

1. **METHODOLOGY**

In this document, represent the ages; they can single ages (i.e 18, 20, 30 …) or group of ages (i.e [1,4], [5,9], …). In any case, their index varies from to . The years are represented by and their index varies from to .

* 1. **Lee and Carter (1992) and Oeppen(2008) models**

The original Lee and Carter (1992) model is presented as follows:

(1)

Where represent mortality rate at age and year , the average of the logarithm of death rate at age , and are represents respectively the mortality index at the year and the deviation in mortality rate due to changes in the index and are estimated using singular value decomposition (SVD) on the matrix

(2)

are the estimation errors and . In order to determine a unique solution of equation (1), Lee and Carter introduced some constraints to their model

and (3)

In CoDa, Oeppen(2008) proposed a model with the expression which is similar to the Lee and Carter original model. But the model forecasts the number of deaths directly instead of the rate of mortality . By using the centred log-ratio transformation

Where is the geometric mean of year of the numerator, Oeppen(2008) proposed the following formulation

(4)

Where is the age-specific geometric mean of over time, and are respectively the age pattern and the time index and are found by using SVD on equation (4).

and

See Oeppen(2008) for more details.

After fitting the model, is written back to the simplex space using the inverse of centered log-ratio

(5)

The operator is a perturbation operator of CoDa framework. It is used to center the matrix without change the constant sum, the operator is used to reverse the operation done by ; and is also a CoDa operator called *closure*, which is used to transform the data into compositional data. All these operators are well described in Bergeron-Boucher et al. (2017).

As in the original Lee and Carter model, forecasts are made by fitting using a time series method.

* 1. **Coherent forecasting**
     1. **Lee and Carter-coherent**

To forecast mortality of different populations in a group coherently, that is to avoid divergence in long run, Li and Lee (2005) proposed an extended model by using a common and specific factor in the original model proposed by Lee and Carter (1992). Their model has been formulated as follows:

(6)

Where is the mortality rate at age and year of the population, represents the average log-mortality at age x in the population , is the common factor found by applying the Lee and Carter(1992) to the average mortality of the whole group. The term is the same term as in Lee and Carter model but specific to each population and is the errors term specific to the population .

Li and Lee (2005) noticed that must approach to a constant over time for the model to work (that is to avoid divergence in mortality rate for each population of the same group over the long run). They also noticed that the model does not guarantee this convergence to a constant and that can make the model fail. Therefore, they suggest to use a random walk without drift or first order auto-regressive AR(1) model with the constant term (Bergeron-Boucher et al., 2017).

* + 1. **CoDa-coherent**

Following the same logic to build a coherent mortality forecasting model for CoDa, Bergeron-Boucher et al. (2017) modified the model proposed by Oeppen (2008). They incorporated a common and specific factor to the model as Li and Lee (2005) did for the original model proposed by Lee and Carter (1992). The CoDa-coherent model proposed by Bergeron-Boucher et al. (2017) is given by

(7)

The formulation is similar to the Lee and Carter coherent model proposed by Li and Lee (2005). Here also, is the geometric mean of the number of deaths at age in the population . The term is common for every populations of the group since it does not depend on . It is found by applying Oeppen(2008) CoDa model to the average of the whole group.

The term represents the perturbation factor specific to every population and is obtained by applying SVD to equation (7). The time series is also forecasted using time series methods that make approach to a constant. As for Li and Lee (2005), the CoDa-coherent model proposed does not guarantee that will convergence.

* 1. **Cohort effect models**
     1. **LC-Cohort of Renshaw and Haberman (2006)**

In order to consider the cohort effect in mortality forecasting, Renshaw and Haberman (2006) extended the model proposed by Lee and Carter (1992) by adding a supplementary factor . Their extended Lee and Carter mortality model with cohort effect is then given by

(8)

The term and represent respectively the cohort effect and its interaction with age. Since is the year and the age, the year of birth is then represented by .

As proposed by Renshaw and Haberman (2006), Oeppen(2008), modelling will be done directly on the number of deaths () instead of on the death rate () proposed by Lee and Carter (1992). Such a modelling is very useful if we want to simulate future cash flows of an annuity for instance (Renshaw and Haberman, 2006). In order to identify a unique solution, the authors imposed the following constraints

and

* + 1. **CODA-Cohort**

Similarly, as Oeppen (2008) did for Lee and Carter model by using CoDa framework we propose to write the CoDa version of Renshaw and Haberman (2006) cohort model as below with a little difference

(10)

Where is the geometric mean of of year ; is the age-specific geometric mean of the number of deaths over time. The remaining parameters are as in Renshaw and Haberman (2006) for . Depending on the values assigned to and it is possible to define sub-categories of the model in (10) similar to those in Renshaw and Haberman (2006).

Considering the expected value part, we can re-write as

(11)

And using the logarithm transformation we have

(12)

Which is the corresponding link function in Renshaw and Haberman (2006) under Poisson assumption.

By considering the expected value again we have

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  | (13) |

Where

Since the centred log-ratio bring data out of simplex space to the Euclidean space, the parameters in (13) can be estimated using the Newton-Raphson deviance minimization method used by Renshaw and Haberman (2006). Under Poisson assumption, the deviance is given by

(14)

Since our main variable of interest is instead of , we will not minimize the deviance of but that of . Therefore, and must be replaced respectively by and in equation (14).

Considering different geometric mean for each year leads to high fluctuations and poor result. This problem is fixed when we consider that every year has the same general geometric mean which is their average over years. Therefore will be taken as the average for all the years in the equation

In general, a parameter is updated using the formula

And under Poisson assumption, parameters updating formula are given as in Renshaw and Haberman (2006)

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Where is equal to if and if .

The algorithm for fitting the CoDa-cohort model is similar to the one proposed by Renshaw and Haberman (2006), is given as follows:

* Step 1: Set starting values
* Step 2: Given and , update and
* Step 3: Given and , update and
* Step 4: Compute

Where is the total number of ages (or group of ages).

As in Renshaw and Haberman (2006), we repeat the cycle form step 2 until converges ().

As for most of the mortality forecasting models based on Lee and Carter (1992), forecasting future number of deaths will be based on a time series models. Only the parameters involving time (i.e and ,) will be predicted. The predicted value of is given by the following

expression:

(15)

The following notations can be found in Renshaw and Haberman (2006)

And

The dot on the top of parameters stands for predicted value.

For and the random walk with drift is the most used model.

* 1. **CoDa-Cohort-Coherent**

In order to forecast multiple population mortality in a coherent way, Lee and Li (2005) and Bergeron-Boucher et al. (2017) have both used a common factor in their model to capture the general pattern of mortality of the whole group of populations. In the CoDa-Cohort-Coherent model presented in this article, as common factor we suggest to use, in addition to the general pattern used by Lee and Li (2005) and Bergeron-Boucher et al. (2017), the cohort effect. The model is presented as following

(16)

Where is the age-specific geometric mean of , all the parameter that does involve form together the common factor and they are found applying the CoDa-Cohort method to the average of the whole group. The errors are represented by . The remaining parameters, and , specific to each population , are found by applying SVD to the following matrix as Bergeron-Boucher et al. (2017) did for CoDa-coherent model

(17)

As in the previous coherent forecasting models, the time series is also forecasted using time series methods that make approach a constant and there is not guarantee that will convergence. As for Li and Lee (2005), the CoDa-cohort-coherent model proposed does not guarantee that convergence.

1. **DATA**

The data used in this study come from the Wold Health Organization (WHO) website (<http://apps.who.int/gho/data/node.main.LIFECOUNTRY?lang=en>). The data provide life table by country from 2000 to 2016. We will choose for our study men mortality data from six African countries: Burkina Faso, Ivory Coast and Mali from West Africa, Kenya, Ethiopia and Somalia from Eastern Africa.

The total number of deaths for some years does not reach 100,000 (the table radius). For such cases, the remaining number of deaths for each year will be added to the number of deaths of the age with the higher number of deaths in that year.

1. **RESULTS**
   1. **Cohort comparison**

In this section, we will compare CoDa cohort model with that of Renshaw and Haberman (2006) model (RH cohort). We will apply those two models to each of the six countries and compare their goodness of fit using their Akaike information criterion (AIC) and Bayesian information criterion (BIC) based on the observed and their respective fitted value. The table 1 contains the AIC and the BIC.

Compared to the RH cohort model, and using the AIC and BIC as model selection criterion, we can notice that CoDa cohort has the lowest value of AIC and BIC for all the six countries. Therefore, we can say that CoDa cohort model fits better these data than the Renshaw and Haberman (2006) cohort model.

**Table 1: Cohort models Akaike information criterion (AIC) and Bayesian information**

**criterion (BIC)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **RH cohort** | | **CoDa cohort** | |
| **Country** | **AIC** | **BIC** | **AIC** | **BIC** |
| Burkina | 3447.254 | 3826.903 | 3360.036 | 3739.685 |
| Ivory Coast | 3604.273 | 3983.922 | 3554.101 | 3933.75 |
| Mali | 3461.602 | 3841.252 | 3380.949 | 3760.598 |
| Kenya | 3695.329 | 4074.978 | 3469.57 | 3849.219 |
| Somalia | 3575.096 | 3954.745 | 3468.503 | 3848.152 |
| Ethiopia | 3546.997 | 3926.646 | 3460.816 | 3840.465 |

Source: WHO (2000 – 2016) – Authors calculations

* 1. **CoDa-Coherent**

In this section, men mortality data for the six countries has been fitted using CoDa-coherent model Bergeron-Boucher et al. (2017) and the CoDa-cohort-coherent presented in this study. We considered two groups of countries based on their proximity criteria and their region in Africa.

The Table 2 contains the Aitchinson distance (AD) to measure the dissimilarity in compositional data analysis (see Bergeron-Boucher et al. (2017) or Aitchison et al. (2000) for more details)

Where is observed composition and the fitted composition.

Table 2 also contains the mean error (ME) which will be used here to measure the fitting bias for the two methods See Bergeron-Boucher et al. (2017) for details.

Considering Aitchinson distance, we notice that CoDa-cohort-coherent model has the lower AD for four countries out of the six. So, CoDa-cohort-coherent model fits these data better than the simple CoDa-coherent model. For ME, both of the CoDa-coherent and CoDa-cohort-coherent have two negative value which indicates that they have the same fitting bias in their fitting method.

**Table 2: Comparison of CoDa coherent and CoDa cohort coherent models based on ME and**

**AD**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **CoDa Coherent** | | **CoDa Cohort Coherent** | |
| **Country** | **ME** | **AD** | **ME** | **AD** |
| Burkina | 3.43E-06 | 0.6198602 | 3.57E-06 | 0.542669 |
| Ivory Coast | -3.24E-06 | 0.6726697 | -2.80E-06 | 0.5836142 |
| Mali | 2.68E-07 | 0.6089474 | 7.64E-07 | 0.5866879 |
| Kenya | 2.67E-05 | 1.259117 | 3.23E-05 | 1.165659 |
| Somalia | -2.92E-05 | 1.538926 | -1.21E-05 | 1.603129 |
| Ethiopia | 9.56E-06 | 0.7801568 | 1.84E-05 | 0.903733 |

Source: WHO (2000 – 2016) – Authors calculations

1. **CONCLUSION AND REMARKS**

The prediction of mortality is a field that continues to grow as the amount of research devoted to it is large. Despite a fairly recent introduction, CoDa is already starting to have a major impact on this area. From the Lee and Carter CoDa version of Oeppen(2008) to CoDa coherent of Bergeron-Boucher et al. (2017), this study continues in the same direction by proposing a Renshaw and Haberman (2006) cohort CoDa version and a CoDa-cohort-coherent.

By using men mortality data of six African countries, it observed that the CoDa-cohort model offers a better fit to the data using AIC and BIC compared to the original RH cohort model. In the case of coherent models, both the simple CoDa-coherent and the CoDa-cohort-coherent model, the same level of bias in their fitting method was observed. But considering the dissimilarity of compositions, the CoDa-cohort-coherent model was the one with the high number of countries with the lowest AD. Therefore, it also indicates better fits for the data than the simple CoDa-coherent. However, some assumptions have been made during the fitting process.

The first is the value of that has been set to the average of geometric mean of all years. That means every year has the same geometric mean of the number of deaths. This assumption was necessary to reduce the fluctuation in order to get a more stable and accurate estimation. As . The second assumption comes from the formulation of the CoDa-cohort-coherent model (equation 16). Indeed, the age-cohort term does not involve . That means every individual with the same year of birth, it doesn’t matter which population he comes from, will have the same cohort effect. That might work fine when the populations are for instance, male and female and the group is a country, a town, etc. But, if the populations are countries, regions, etc. each of these populations are likely to have cohort effect specific to each of them. This can be an issue for further research.

**References**

Bergeron-Boucher, M.-P., Canudas-romo, V., and Vaupel, J. W. (2017). Coherent Forecasts of

Mortality with Compositional Data Analysis. *Demographic Research* 37(August):527–566.

Lee, R. D. and Carter, L. R. (1992). Modeling and Forecasting U.S. Mortality Modeling and

Forecasting U.S. Mortality*. Journal of the American Statistical Association*.

Li, N. and Lee, R. (2005). Forecasts for a Coherent Mortality an Extension of Group of

Population: The Lee-Carter method. *Demography* 42(3):575–594.

Oeppen, J. (2008). Coherent forecasting of Multiple-decrement Life Tables: a test using Japanese

cause of death data. (1992). *Paper presented at the European Population Conference,*

*Barcelona, Spain, July 9–12, 2008*.

Renshaw, A. E. and Haberman, S. (2006). A Cohort-Based Extension to the Lee – Carter

model for Mortality Reduction Factors. *Insurance: Mathematics and Economics*

38:556–570.