Forecasting causes of death using compositional data analysis - the case of cancer deaths

Søren Kjærgaard, Yunus Emre Ergemen, Jim Oeppen, Malene Kallestrup-Lamb, and Rune Lindahl-Jacobsen

Center on Population Dynamics (CPOP), University of Southern Denmark

September 12, 2018
Cancer deaths for Dutch males by age

Forecasting cause of deaths

September 12, 2018
Proportion of deaths for Dutch males from 1957 to 2014

Forecasting cause of deaths

September 12, 2018
Objective and data

- Forecast the number of cancer deaths

Useful for health care planning

If we also can forecast relative risk, we can forecast incidence rates

Data for French and Dutch populations
Objective and data

- Forecast the number of cancer deaths
- Useful for health care planning
Objective and data

- Forecast the number of cancer deaths
- Useful for health care planning
- If we also can forecast relative risk, we can forecast incidence rates
Objective and data

- Forecast the number of cancer deaths
- Useful for health care planning
- If we also can forecast relative risk, we can forecast incidence rates
- Data for French and Dutch populations
Life table deaths are composition data

- We use life table deaths \((d_{x,t,i})\)
Life table deaths are composition data

- We use life table deaths \((d_{x,t,i})\)
- Age \(x\), time \(t\), and cause \(i\)
Life table deaths are composition data

- We use life table deaths \((d_{x,t,i})\)
- Age \(x\), time \(t\), and cause \(i\)
- \(\sum_i \sum_x d_{x,i} = 1\)
Life table deaths are composition data

- We use life table deaths \((d_{x,t,i})\)
- Age \(x\), time \(t\), and cause \(i\)
- \(\sum_i \sum_x d_{x,i} = 1\)
- Thus, life table deaths are compositional data
Life table deaths are composition data

- We use life table deaths \((d_{x,t,i})\)
- Age \(x\), time \(t\), and cause \(i\)
- \(\sum_i \sum_x d_{x,i} = 1\)
- Thus, life table deaths are compositional data
- Can be problematic to use standard statistical methods as they are defined for real values
Oeppen (2008) suggests to use a compositional data (CoDa) model outside a Lee-Carter model.
• Oeppen (2008) suggests to use a compositional data (CoDa) model outside a Lee-Carter model
• Causes are stacked horizontally
Oeppen (2008) suggests to use a compositional data (CoDa) model outside a Lee-Carter model.

Causes are stacked horizontally.

\[
clr(d_{t,x,i} \Theta \alpha_{x,i}) = \beta^{1}_{x,i} k_{t}^{1} + \beta^{2}_{x,i} k_{t}^{2} + \ldots + \beta^{p}_{x,i} k_{t}^{p} + \epsilon_{t,x,i}, \quad (1)
\]
Common trend CoDa model (CT-CoDa) Oeppen (2008)

Forecasting cause of deaths
CT-CoDa model limitations

- After centring, causes are weighted equally
CT-CoDa model limitations

- After centring, causes are weighted equally
- $\beta_{x,i}$ is assumed to be stable over time
CT-CoDa model limitations

- After centring, causes are weighted equally
- $\beta_{x,i}$ is assumed to be stable over time
- Variation is decomposed when common for all causes
CT-CoDa model limitations

- After centring, causes are weighted equally
- $\beta_{x,i}$ is assumed to be stable over time
- Variation is decomposed when common for all causes
- Only one time trend is assumed for each rank approximation for all causes
Centered deaths for Dutch males in selected years

Death distribution

Centred death distribution

Forecasting cause of deaths

September 12, 2018
2 step CoDa model (2S-CoDa)

\[ w_{age}^{x,i} = \frac{\bar{d}_{x,i}}{\sum_{x=1}^{\omega} \sum_{i=1}^{K} \bar{d}_{x,i}} \]

\[ w_{time}^t = \rho \cdot (1 - \rho)^{(T-t)} \]

\[ clr(d_{t,x,i} \ominus \alpha_{x,i}) = \beta_{x,i}^J k_t^J + \beta_{x,i}^I k_t^I + \epsilon_{t,x,i} \]

- Age and cause specific weights
- Time weight
- Decomposing of cause specific variation
VECM CoDa model (VECM-CoDa)

\[
clr(d_{t,x,i} \ominus \alpha_{x,i}) = \beta_{x,i}^1 k_{t,i}^1 + \ldots + \beta_{x,i}^p k_{t,i}^p + \epsilon_{t,x,i},
\]

- Allows for cause specific time trends
- Dependence is modelled by determining stationary relationships between the time trends
VECM CoDa model (VECM-CoDa)

\[ \Delta k_t = \Pi k_{t-1} + \sum_{j=1}^{\Gamma} \Gamma_j \Delta x_{t-j} + B + \epsilon_t \]

- \( \Pi \) has rank zero meaning there are no long run relationships among the series, but the series are non-stationary.
- \( \Pi \) has full rank which means that all of the series are stationary.
- \( \Pi \) has reduced rank, \( r > 0 \), thus there exist both stationary and non-stationary series and \( r \) stable long run relationships exist.
Forecasting cause of deaths

September 12, 2018
Explained variation in the 2S-CoDa model

Forecasting cause of deaths
Weights in the 2S-CoDa model

Forecasting cause of deaths
15 years out-of-sample forecasts for Dutch males

- Deaths
- Observed
- CT-CoDa
- 2S-CoDa
- VECM-CoDa
- LC

Forecasting cause of deaths

September 12, 2018
Table: 20 years out-of-sample forecast error with rolling origin, for French and Dutch populations

<table>
<thead>
<tr>
<th>Model</th>
<th>FRA females</th>
<th>FRA males</th>
<th>NLD females</th>
<th>NLD males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE measured in life table deaths</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT-CoDa</td>
<td>105.4</td>
<td>292.3</td>
<td>140.03</td>
<td>316.9</td>
</tr>
<tr>
<td>2S-CoDa</td>
<td><strong>90.9</strong>*</td>
<td><strong>217.2</strong>*</td>
<td>168.55</td>
<td><strong>259.1</strong>*</td>
</tr>
<tr>
<td>VECM-CoDa</td>
<td>114.6</td>
<td>369.4</td>
<td><strong>108.13</strong></td>
<td>391.6*</td>
</tr>
<tr>
<td>LC</td>
<td><strong>99.6</strong>*</td>
<td><strong>263.9</strong>*</td>
<td>153.56</td>
<td>484.3</td>
</tr>
<tr>
<td></td>
<td>RMSE measured in deaths rates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT-CoDa</td>
<td>0.00076</td>
<td>0.00320</td>
<td>0.00631</td>
<td>0.01366</td>
</tr>
<tr>
<td>2S-CoDa</td>
<td><strong>0.00070</strong>*</td>
<td><strong>0.00239</strong>*</td>
<td><strong>0.00586</strong></td>
<td><strong>0.01349</strong>*</td>
</tr>
<tr>
<td>VECM-CoDa</td>
<td>0.00085</td>
<td>0.00570</td>
<td>0.00634</td>
<td>0.01490*</td>
</tr>
<tr>
<td>LC</td>
<td><strong>0.00056</strong>*</td>
<td>0.00324*</td>
<td>0.00654</td>
<td>0.01571</td>
</tr>
</tbody>
</table>

* indicates that the model is significantly different from the CT-CoDa model on a 5% significant level using the Clark-West test.
Selection of causes when forecasting cancer

Questions: Are all causes needed for an accurate forecast of total number of cancer deaths?

**Table:** Elastic net results for Dutch males

<table>
<thead>
<tr>
<th></th>
<th>25-29</th>
<th>30-34</th>
<th>35-39</th>
<th>40-44</th>
<th>45-49</th>
<th>50-54</th>
<th>55-59</th>
<th>60-64</th>
<th>65-69</th>
<th>70-74</th>
<th>75-79</th>
<th>80-84</th>
<th>85-89</th>
<th>90-94</th>
<th>95+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infectious</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.0380</td>
<td>-0.1954</td>
<td>-0.1139</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0457</td>
<td>0</td>
</tr>
<tr>
<td>Endocrine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.0276</td>
<td>-0.0633</td>
<td>-0.0534</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0442</td>
<td>0.0684</td>
</tr>
<tr>
<td>Mental</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.0108</td>
<td>-0.0552</td>
<td>-0.0419</td>
<td>-0.0345</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nervous</td>
<td>0.0115</td>
<td>0.1120</td>
<td>0.0735</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1333</td>
<td>0.2009</td>
<td>0.1855</td>
<td>0.1771</td>
<td>0.0459</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Circular</td>
<td>0.0508</td>
<td>0.1887</td>
<td>0.2542</td>
<td>0.3330</td>
<td>0.3519</td>
<td>0.3664</td>
<td>0.3340</td>
<td>0.2992</td>
<td>0.1834</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1168</td>
<td></td>
</tr>
<tr>
<td>Respiratory</td>
<td>0.1242</td>
<td>0.1138</td>
<td>0.0099</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0092</td>
<td>0.3434</td>
<td>0.3137</td>
<td>0.1098</td>
<td>0.0204</td>
<td>0.2281</td>
<td>0.5073</td>
<td></td>
</tr>
<tr>
<td>Digestive</td>
<td>0.1360</td>
<td>0.0443</td>
<td>0</td>
<td>0.0158</td>
<td>0.0082</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2411</td>
<td>0.3140</td>
<td>0.0040</td>
</tr>
<tr>
<td>Other</td>
<td>0.0903</td>
<td>0.2707</td>
<td>0.2276</td>
<td>0.1293</td>
<td>0.0836</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-0.1089</td>
<td>-0.0429</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.3487</td>
</tr>
<tr>
<td>External</td>
<td>0.3343</td>
<td>0</td>
<td>0.1706</td>
<td>0.0888</td>
<td>0.0064</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.77</td>
<td>0.87</td>
<td>0.83</td>
<td>0.89</td>
<td>0.91</td>
<td>0.93</td>
<td>0.94</td>
<td>0.90</td>
<td>0.80</td>
<td>0.83</td>
<td>0.80</td>
<td>0.82</td>
<td>0.90</td>
<td>0.90</td>
<td>0.85</td>
</tr>
</tbody>
</table>
## Forecasting errors when dropping causes for Dutch males

<table>
<thead>
<tr>
<th>Model</th>
<th>All included</th>
<th>Drop COD 10</th>
<th>Drop COD 8, 10</th>
<th>Drop COD 6, 8, 10</th>
<th>Drop COD 4, 6, 8, 10</th>
<th>Drop COD 3, 4, 6, 8, 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT-CoDa</td>
<td>0.1351</td>
<td>0.1318</td>
<td>0.1293</td>
<td>0.2020</td>
<td>0.2443</td>
<td>0.2545</td>
</tr>
<tr>
<td>2S-CoDa</td>
<td>0.1068</td>
<td>0.0998</td>
<td>0.0970</td>
<td>0.1493</td>
<td>0.1629</td>
<td>0.1585</td>
</tr>
<tr>
<td>VECM-CoDa</td>
<td>0.1691</td>
<td>0.1557</td>
<td>0.1651</td>
<td>0.1819</td>
<td>0.1955</td>
<td>0.2215</td>
</tr>
</tbody>
</table>

COD1 (Infectious diseases), COD2 (Cancer), COD3 (Endocrine diseases), COD4 (Mental diseases), COD5 (Nervous diseases), COD6 (Circular diseases), COD7 (Respiratory diseases), COD8 (Digestive diseases), COD9 (Other diseases), COD10 (External)
Conclusions

- Introducing weights and decomposition of cause specific variation can improve the model suggested by Oeppen (2008)
- Allowing for multiple time trends did not improve the forecast accuracy
- Dropping causes can improve the forecast performance but a forecast-bias is introduced because of dependence among the causes.