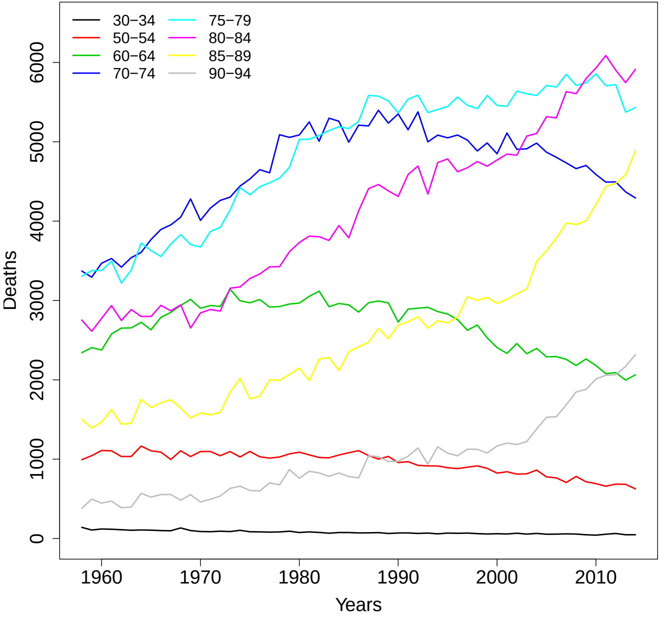
Forecasting Causes of Death using Compositional Data Analysis - the case of Cancer Deaths

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

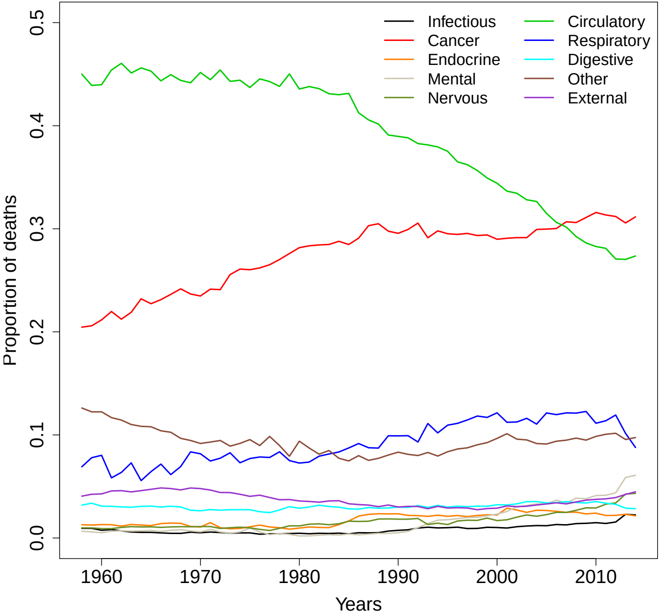
Center on Population Dynamics (CPOP), University of Southern Denmark

September 12, 2018

Cancer deaths for Dutch males by age



Proportion of deaths for Dutch males from 1957 to 2014



Objective and data

- Forecast the number of cancer deaths

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

- 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
- $\sum_i \sum_x d_{x,i} = 1$

- 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

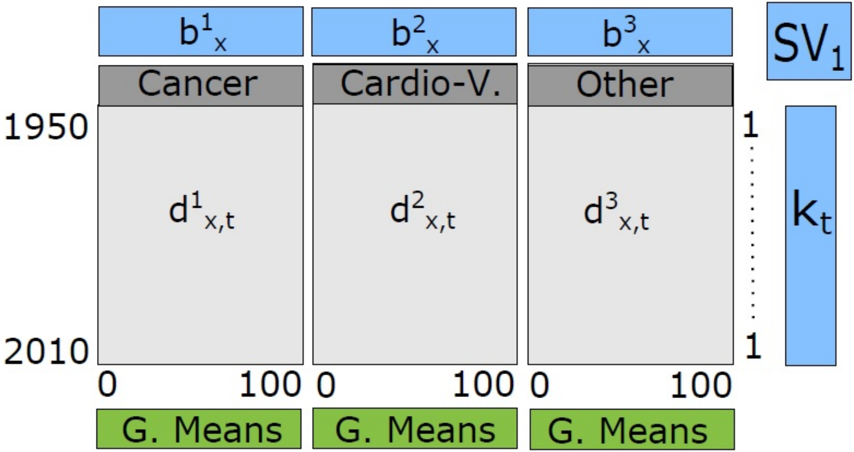
- Oeppen (2008) suggests to use a compositional data (CoDa) model outside a Lee-Carter model

- 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

$$clr(d_{t,x,i} \ominus \alpha_{x,i}) = \beta^1_{x,i} k^1_t + \beta^2_{x,i} k^2_t + ... + \beta^p_{x,i} k^p_t + \epsilon_{t,x,i}, \quad (1)$$

Commmon trend CoDa model (CT-CoDa) Oeppen (2008)



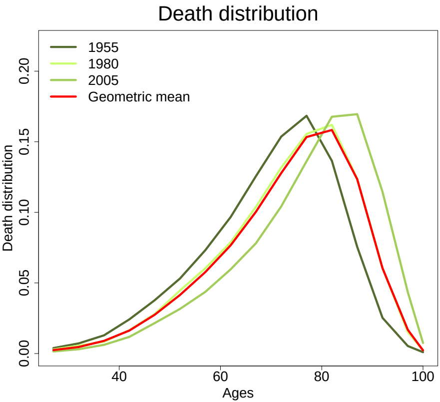
- After centring, causes are weighted equally

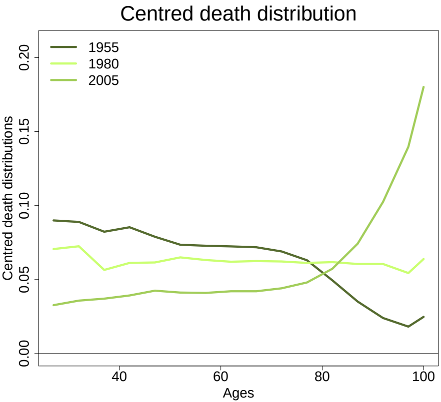
- After centring, causes are weighted equally
- $\beta_{x,i}$ is assumed to be stable over time

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

- 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





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

$$w_t^{time} = \rho \cdot (1-\rho)^{(T-t)}$$

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

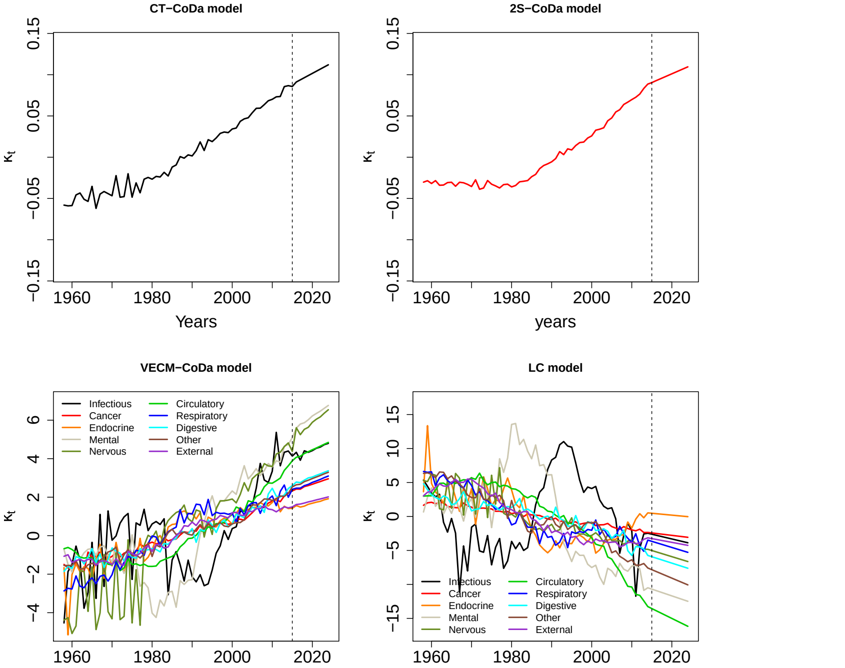
- Age and cause specific weights
- Time weight
- Decomposing of cause specific variation

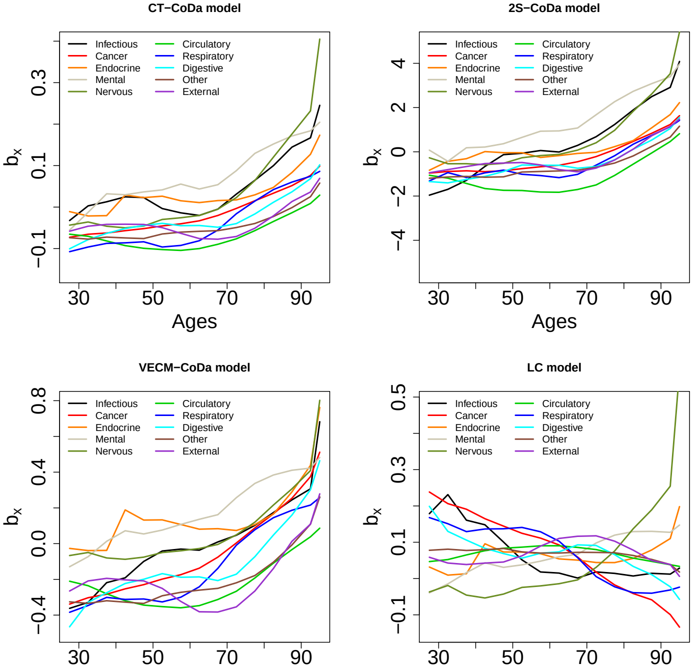
$$clr(d_{t,x,i} \ominus \alpha_{x,i}) = \beta^1_{x,i} k^1_{t,i} + ... + \beta^p_{x,i} k^p_{t,i} + \epsilon_{t,x,i},$$

- Allows for cause specific time trends
- Dependence is modelled by determining stationary relationships between the time trends

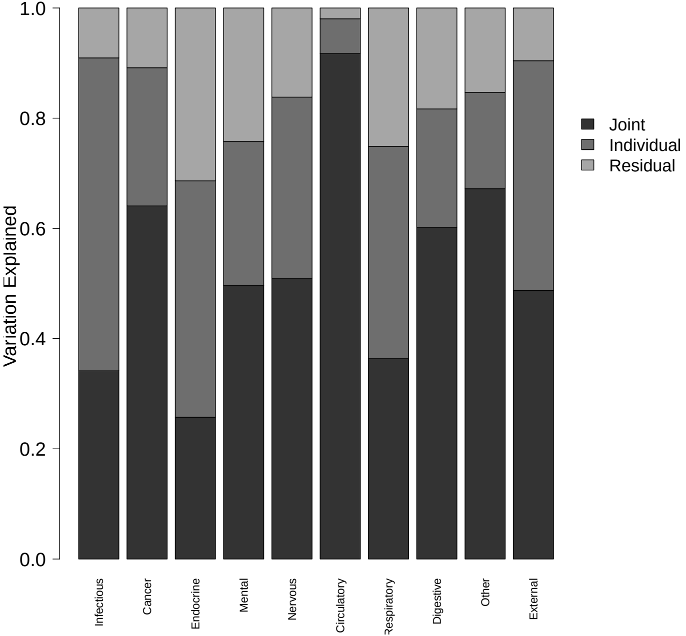
$$\Delta k_t = \Pi k_{t-1} + \sum_{j=1} \Gamma_j \Delta x_{t-j} + B + \epsilon_t$$

- Π has rank zero meaning there are no long run relationships among the series, but the series are non-stationary.
- Π has full rank which means that all of the series are stationary.
- Π has reduced rank, $r > 0$, thus there exist both stationary and non-stationary series and r stable long run relationships exist.

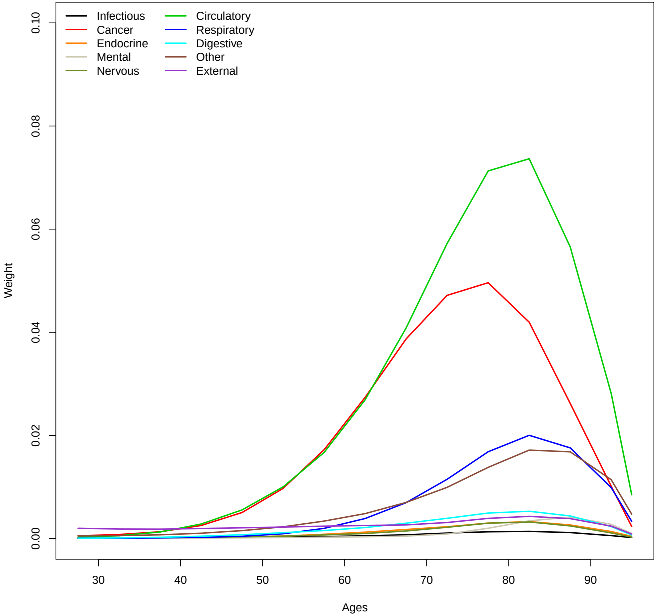

CT−CoDa model
2S−CoDa model
VECM−CoDa model
LC model
κ_t
Years
years
Infectious
Cancer
Endocrine
Mental
Nervous
Circulatory
Respiratory
Digestive
Other
External


CT−CoDa model
2S−CoDa model
VECM−CoDa model
LC model
Infectious
Cancer
Endocrine
Mental
Nervous
Circulatory
Respiratory
Digestive
Other
External
b_x
Ages

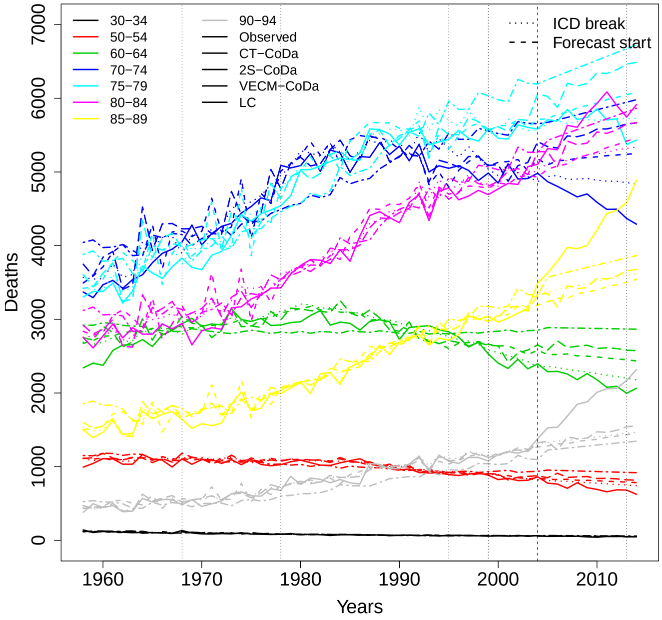
Explained variation in the 2S-CoDa model



Weights in the 2S-CoDa model



15 years out-of-sample forecasts for Dutch males



Proportions of deaths for Dutch males

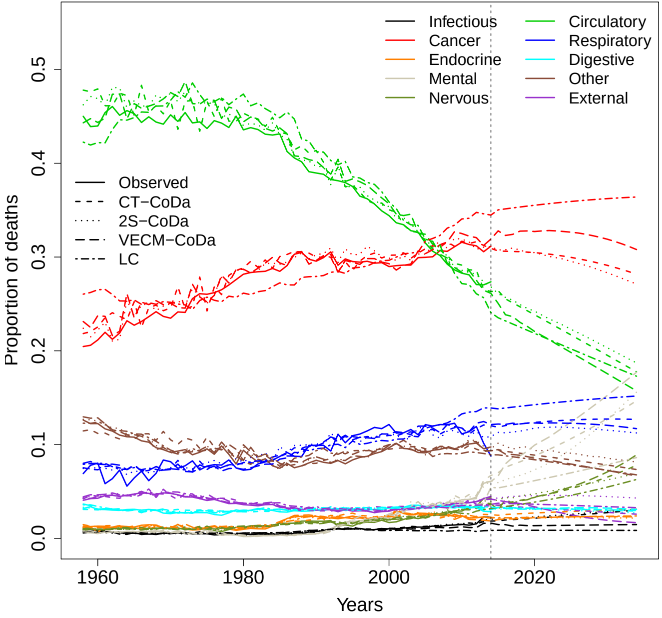


Table: 20 years out-of-sample forecast error with rolling origin, for French and Dutch populations

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

* indicates that the model is significantly different from the CT-CoDa model on a 5% significant level using the Clark-West test.

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

Table: Elastic net results for Dutch males

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

Forecasting errors when dropping causes for Dutch males

Model	All included	Drop COD 10	Drop COD 8, 10	Drop COD 6, 8, 10	Drop COD 4, 6, 8, 10	Drop COD 3, 4, 6, 8, 10
CT-CoDa	0.1351	0.1318	0.1293	0.2020	0.2443	0.2545
2S-CoDa	0.1068	0.0998	0.0970	0.1493	0.1629	0.1585
VECM-CoDa	0.1691	0.1557	0.1651	0.1819	0.1955	0.2215

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.