Title: Robustness and bias of excess death estimates in 2020 under varying model specifications

Authors: Jonas Schöley[1]

Affiliations:

[1] Interdisciplinary Centre on Population Dynamics, University of Southern Denmark; Odense 5000, Denmark

Corresponding authors:

Jonas Schöley jschoeley@health.sdu.dk

Keywords: excess deaths; COVID-19; Serfling model; cross-validation; robustness

Abstract: Various procedures are in use to calculate excess deaths during the ongoing COVID-19 pandemic. Using weekly death counts from 20 European countries, we evaluate the robustness of excess death estimates to the choice of model for expected deaths and perform a cross-validation analysis to assess the error and bias in each model's predicted death counts. We find that the different models produce very similar patterns of weekly excess deaths but disagree substantially on the level of excess. The country ranking of percent excess death in 2020 is sensitive to the exclusion of an exposure variable but otherwise stable across models that do account for population structure. The five-year average death rate model consistently produces the lowest excess death estimates, whereas high excess deaths are produced by the popular five-year average death count and Euromomo models. Cross-validation revealed these estimates to be biased under a causal interpretation of "expected deaths had COVID-19 not happened."

Introduction

The concept of "excess deaths" became widely known in 2020. Early in the year, international newspapers reported weekly death counts that were notably elevated compared to previous years [25, 28]. In a situation where tests for COVID-19 were scarce and the underlying cause of death often hard to determine, excess deaths proved to be a rational means of monitoring the direct and indirect mortality impact of the ongoing COVID-19 pandemic with minimal data requirements [14]. Since the start of the pandemic, the literature on excess deaths has been growing fast and already features international comparisons and rankings owing to the readily available data on death counts for many countries [13, 11, 20, 3]. However, despite the prominence and relevance of the measure, the excess death methodology is far from standardized with different definitions of excess deaths and different models used to estimate an expected death counts against which the excess is judged. This eclectic state of the art begs the questions on robustness and bias of excess death estimates and cross-country comparisons under various analysis strategies.

Given data from 20 European countries, we will assess how excess death measures in the year 2020 vary under different modeling choices for the expected number of deaths. Furthermore, we test the models for bias in their predicted weekly and annual death counts using a time-series cross-validation setup. While the lack of consensus extends beyond the model specification to the very definition of excess deaths, we will focus entirely on the baseline model's impact and define excess as observed minus expected, allowing the measure to be negative.

As the literature on excess deaths in 2020 is dominated by simple average or regression-based approaches, we will restrict our attention to this class of models, focusing on the effect of different specifications. Models based upon multi-year averages are prominently employed by newspapers and statistical offices and featured in the academic literature [4, 15, 16, 18]. While simple and easy to communicate, these procedures do not adjust for time trends in the death counts or rates. This issue is addressed by the Serfling model [22], which in various specifications has been applied to quantify covid related deaths [2, 26, 27, 8, 11]. Generalized additive models relax the strict specifications of the Serfling model and allow for smooth long-term and seasonal effects [1, 21]. These models may be further elaborated via the inclusion of temperature effects, autoregressive residuals, adjustments for bank holidays, and Bayesian inference [13].

Having chosen 10 model specifications reflecting the heterogeneity in the literature (Table 1), we probe each model for a tendency to produce high or low estimate of expected and conversely excess deaths, both weekly and total. We ask if the ranking of European countries along the total percentage of excess deaths in 2020 differs under different baseline models, and we identify those countries where there is disagreement about the existence of excess deaths among the models.

When excess death numbers vary by models, the question arises which model to trust. In line with [13] and [17] we argue that the challenge of estimating the expected number of weekly deaths in the year 2020, given that COVID had not occurred, is a classical forecasting challenge, and thus, the models can be validated by testing their predictions on past data. To that end, we construct a cross-validation challenge, where each model is fitted on seven years of weekly death counts and is then tasked to predict the next 52 weeks from that series. The error and bias of any model can then be estimated by comparing the model predictions with the observed data. If a model's predictions tended to over/underestimate weekly deaths in the past, we argue that this model also over/underestimates the expected deaths in 2020 under the counterfactual no-COVID scenario, thereby biasing excess deaths.

Data and Methods

Raw and derived data

To replicate the more elaborate expected death regression models, we collected cross-country data on weekly death counts and population exposure by age and sex, along with information on the timing of public holidays and weekly population-weighted temperature anomalies.

Age and sex-specific weekly death counts by country were sourced from the Short-term Mortality Fluctuations (STMF) database [12, 19]. We selected all European countries with data going back to at least 2007. As we are using the Mean Absolute Percentage Error (MAPE) measure to report on prediction error, we needed to ensure strictly positive weekly death counts. We thus excluded Estonia, Iceland, and Luxembourg from the analysis and summed deaths into age categories 0 to 65, 65 to 75, 75 to 85, and 85+.

Person-weeks of population exposure by country, sex, and age were derived from mid-year population estimates from the Human Mortality Database [10]. We summed these estimates into the target age groups and then interpolated over the weeks of a year by fitting a cubic spline and extrapolating linearly to the end of 2020. Integrating the spline over single weeks yielded person-weeks of exposure.

Population weighted weekly temperature anomalies were calculated from global gridded temperature [6] and population data [5]. First, we calculated the average weekly temperature per grid cell from the daily measurements and then weighted these measurements by the population size in the same grid cell. We then selected all grid-cells overlapping with the area of a given country and calculated the country's population-weighted average weekly temperature. By averaging the weekly temperatures over multiple years, we established an expected temperature. A country's weekly population-weighted temperature anomalies were then derived by subtracting this long-term trend from the weekly temperature in a given year.

We used the "Nager.Date" software [9] to determine for each week, year, and country the occurrence and type of a public holiday, distinguishing between Christmas, Easter, New Year, and "other" types of public holidays.

Model choice

We compare the 10 different models listed in Table 1 in terms of their predicted death counts and their prediction error on historical data. The models are chosen to represent the range of approaches in the literature on excess deaths estimation since the start of the pandemic, from very simple to elaborate, but we do not claim to have exactly replicated any published analysis. All models are either fitted separately for each combination of age and sex or feature interaction terms to the same effect. Further details are given in the supplementary materials. We chose not to include a flu-epidemic variable in any of our models due to endogeneity. Mitigation measures put in place to combat the spread of the Coronavirus also have an impact on the spread of other viruses [7, 23]. Therefore, if the expected deaths are to be interpreted as "deaths if COVID had not happened", including post-covid influenza observations is not admissible.

Table 1: Models for weekly expected deaths.

Model	Description	References
Weekly averages (AVG5c) 5 year average death counts (AVG5r) 5 year average death rates	simple mean over the weekly deaths counts or death rates in the preceding years	[18], [24], [4]
Serfling Model (SRFc) without exposures (SRFcem) Euromomo style, i.e. no exposures, no bank holiday coefficient, fitted over 5 years on weeks without flu-activity (SRFr) with exposures	Poisson regression on death counts with log-linear long term trend and Fourier-term seasonality	[2], [26], [27], [8], [11]
Generalized Additive Model (GAMr) without temperature anomaly predictor (GAMrt) with temperature anomaly predictor as smoothly varying coefficient over week of year	Poisson regression on death counts with log-linear long term trend and cyclical spline seasonality	[1], [21]
Latent Gaussian Model (LGMr) without temperature anomaly predictor (LRMrt) with temperature anomaly predictor as varying coefficient over week of year implemented as cyclic random walk (LGMrt2) same as LRMrt with order 2 autoregressive long term trend and temperature anomaly random walk	Bayesian Poisson regression on death counts with autoregressive long term trend and time-varying seasonality	[13]

Excess death measures

We define excess deaths D^* in a given week w and stratum j as $D^*_{wj} = D_{wj} - \widehat{D}_{wj}$, where $\widehat{D}$ are the expected deaths under a given model. This definition as a model residual allows capturing eventual mortality displacement effects where a positive excess in part of the year can be followed by negative excess in later weeks. The total number of excess deaths over period $[a, b]$ are given by $D^*_{[a,b],j} = D_{[a,b],j} - \widehat{D}_{[a,b],j}$, where $D_{[a,b],j} = \sum_{w=a}^{b} D_{wj}$ and $\widehat{D}_{[a,b],j} = \sum_{w=a}^{b} \widehat{D}_{wj}$. Following a common strategy, expected and observed deaths may be summed over strata to estimate the excess at a higher aggregation level [2, 1, 13, 26].

To facilitate comparisons between countries, we report the weekly percentage increase of observed deaths over expected deaths, $P_{wj} = \frac{D^*_{wj}}{\widehat{D}_{wj}}$, the so-called "P-score". For a sequence of weeks $[a, b]$

we calculate the cumulative P-score from the total excess and expected deaths over that period, $P_{[a,b],j} = \frac{D^*_{[a,b],j}}{\widehat{D}_{[a,b],j}}$.
Any prediction intervals around the quantities above are derived from repeated samples of expected deaths from a Poisson distribution with rate $\widehat{D}_{wj}$. The resulting uncertainty statements should be understood as lower-bound estimates as the Poisson sampling, while easily implemented for any algorithm which predicts death counts, ignores overdispersion in the outcome distribution and uncertainty around model parameters.

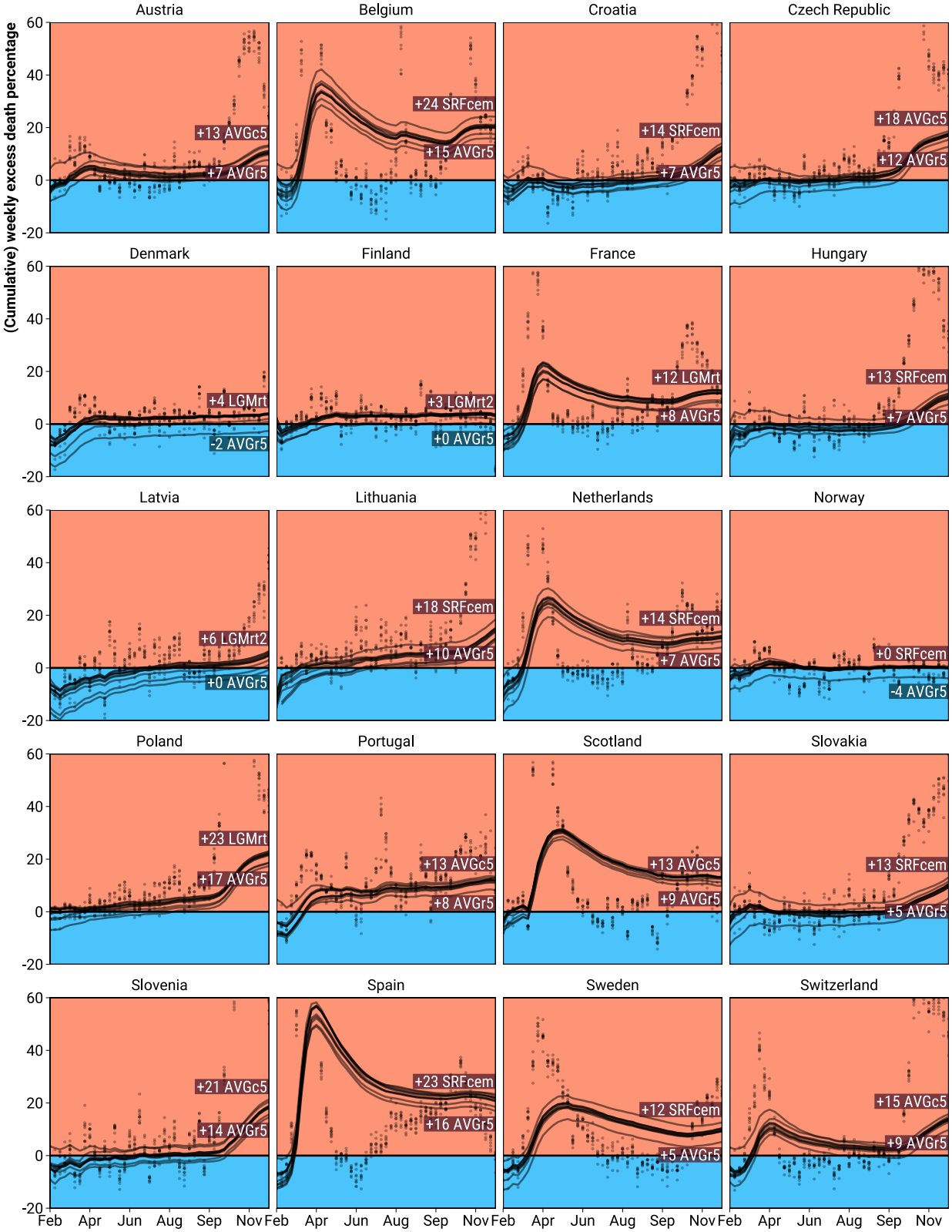
Cross-validation

Time series of weekly death counts and co-variates were split into five cross-validation sets. The training period starts at week 27 of a year and ends after seven fully observed summer to summer seasonal cycles with the beginning of week 8 of a year. After fitting the model, weekly death counts were predicted for 45 weeks following week 8 of the last year in the training set. The starting years for the five cross-validation series were 2007 to 2011. This particular cross-validation setup has been chosen to mirror the task of predicting weekly deaths in 2020, given years of data observed until shortly before the outbreak of the pandemic.

Results

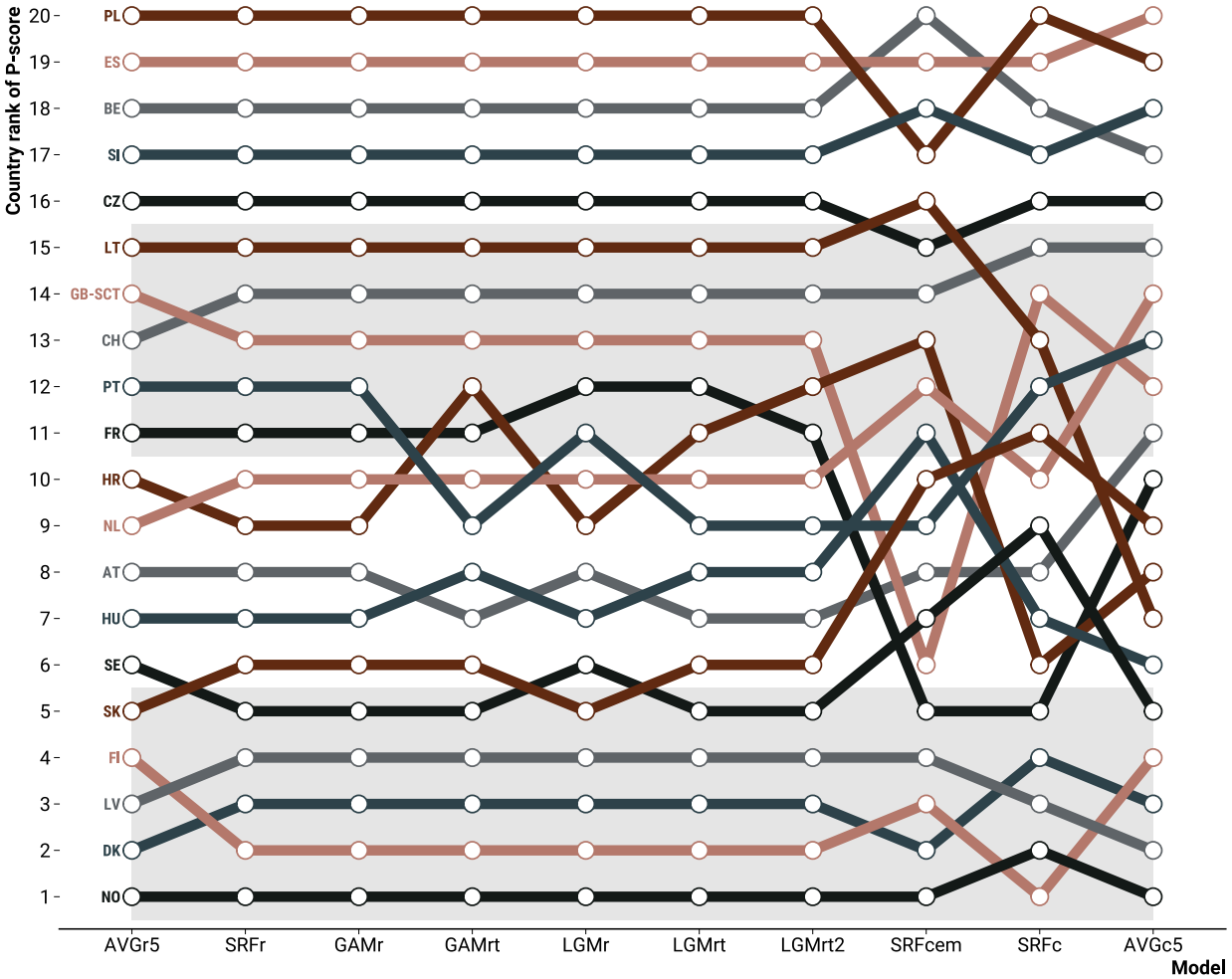
Models differed in their estimated level of excess deaths while generally agreeing on the weekly pattern of excess (Figure 1). On the country level, the range of weekly and annual percent excess deaths across models was between 5 and 10%. The average death rate model produced the lowest excess death estimate for all countries, whereas the maximum varied, with the Euromomo style Serfling and the average death count model producing the highest excess in 15 out of 20 countries. Taking Belgium as an example of a region with pronounced mortality, central estimates of excess deaths by the end of 2020 range from 14,968 (.95 PI: 14,472–15,510) under the average death rate model to 21,485 (.95 PI: 21,068–21,851) under a Euromomo style Serfling regression. Consequently, models may disagree regarding the existence of significantly elevated mortality in countries with comparatively few registered COVID deaths. We found such disagreement in Denmark, Finland, and Latvia, where annual P-scores were not significantly elevated under the average death rate model. Norway is the only country under consideration where all models agreed that deaths in 2020 had not been significantly elevated.

Figure 1: Weekly and cumulative excess death percentage as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions. *Note: Points indicate weekly excess percentages. Cumulative excess percentages are indicated by curves and derived from cumulated observed and expected deaths. The labels refer to the percent excess over the entire analysis period.*



The ranking of countries along the percent excess deaths in 2020 was largely stable across those models which controlled for population structure (AVGr5, SRFr, GAM*, LGM*) but fluctuated once an exposure variable was excluded (Figure 2). All models agreed on the four countries with the highest (Poland, Spain, Belgium, Slovenia) and lowest (Finland, Latvia, Denmark, Norway) annual percent excess. There was substantial disagreement among the exposure-free models (SRFcem, SRFc, AVGc5) regarding the middle ranks with countries moving between quartiles. France and Lithuania are two extreme examples where the European annual P-score rank stretched across three quartiles depending on the model. Within age and sex strata, we observed even more drastic changes in rank across models without exposure variable, whereas the rate-based models largely agreed in their predicted country rankings (Figure S.11).

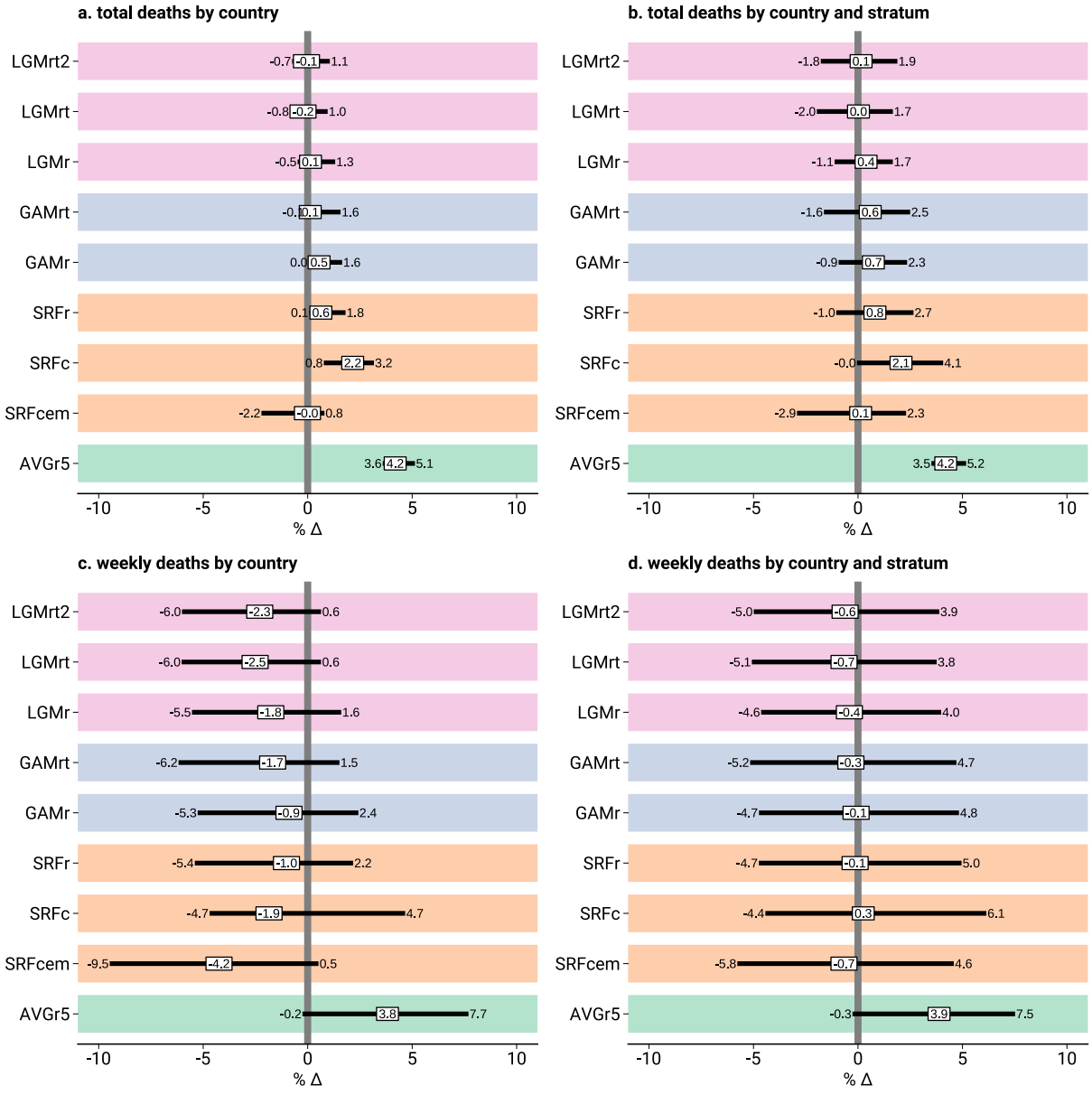
Figure 2: Country ranking of excess death percentage during the year 2020 weeks 8 through 52 under 10 different models.



In Figure 3a we compared the 5-year average death count model (AVGc5, the reference) to various other models concerning the expected number of deaths in 2020. The Euromomo Serfling model (SRFcem) and the reference tended to produce the lowest expected death estimates and thus the highest excess death counts, whereas the 5-year average death rate model (AVGr5) gravitated towards the highest expected death counts, with predictions being higher by 3.6 to 5.1% compared to the reference for half of the countries. The regression models with exposure offset (SRFr, GAM*, LGM*) estimated baseline deaths close to the reference model if somewhat higher, whereas the Serfling model without exposures (SRFc) tended towards substantially higher expected deaths.

The Latent Gaussian Models (LGM*) had the highest agreement with the reference, e.g., for half of the countries, the sophisticated model LGMrt2 produced estimates within -0.7 to 1.1% of the basic average death count prediction. Similar results were found for predictions of total deaths by sex and age strata within a country, albeit with higher variability of the percentage differences for any given model (Figure 3b).

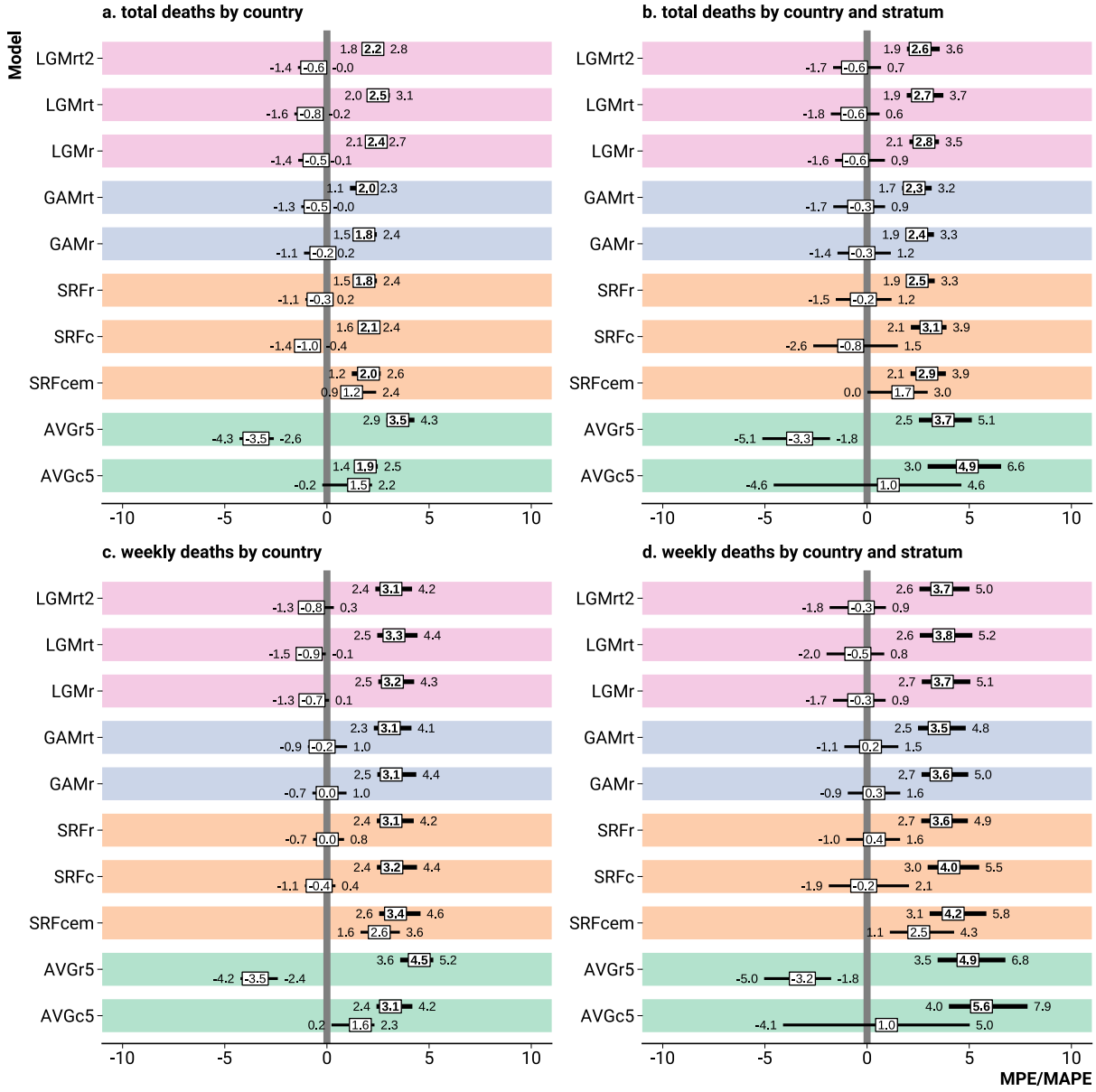
Figure 3: Percent differences of predicted death counts from various models against the 5-year average weekly death prediction. *Note: The prediction differences were summarised across countries and, where applicable, weeks and strata by the 0.25, 0.5, and 0.75 quantiles.*



Percentage differences of predicted weekly death counts against the reference model are shown in Figures 3c and d. Again, the 5-year average death rate model tended to produce the highest estimates, with half of the weekly predicted counts being at least 3.8% above those from the reference. The regression models, on average, predicted fewer deaths than reference, with the Euromomo Serfling model featuring especially low weekly expected deaths. Note, however, that

these percentage differences varied considerably over countries and strata due to the high variability of the weekly predictions.

Figure 4: Distribution of test set errors and bias for predicted death counts. *Note: The MAPE (bold) and the MPE were summarised across countries and, where applicable, weeks and sex*age strata by the 0.25, 0.5, and 0.75 quantiles.*



The cross-validation study confirmed the substantial biases of the Euromomo Serfling regression and the 5-year average death rate and count models (Figure 4a–d). While the average rate model (AVGr5) consistently displayed the highest propensity among all models to overestimate deaths on four different prediction tasks, the average count (AVGc5) and the Serfling Euromomo models (SRFcem) underestimated deaths. Regression models with exposure offset were comparatively unbiased.

When predicting country-level total deaths over weeks 8 through 52 of a year, the median MAPE on the test set ranged from 1.8% for the best performing regression models (SRFr, GAMr) to 3.5%

for the average death rate model. Notably, the MAPE from the 5-year average death count model was comparable to the error of the best-performing regression models when predicting annual or weekly deaths on the country level (Figures 4a and c). However, when predicting deaths by sex and age strata, the regression models performed substantially better than either the average rate or count model, with little difference between the various regression specifications (Figures 4b and d). Cross-validation errors for all models increased by approximately 1 to 2 percentage points going from annual to weekly predictions.

Discussion

[To be continued]

References

[1] J. M. Aburto, R. Kashyap, J. Schöley, C. Angus, J. Ermisch, M. C. Mills, and J. B. Dowd. Estimating the burden of the COVID-19 pandemic on mortality, life expectancy and lifespan inequality in england and wales: a population-level analysis. *Journal of Epidemiology and Community Health*, 0:1–6, jan 2021. doi:10.1136/jech-2020-215505.

[2] S. Barnard, S. Fox, A. Baker, P. Burton, P. Goldblatt, and J. Fitzpatrick. Excess mortality in england. methodology for the weekly reports. Technical Report GW-138, Public Health England, 2020.

[3] A. Bilinski and E. J. Emanuel. COVID-19 and excess all-cause mortality in the US and 18 comparison countries. *JAMA*, 324(20):2100, nov 2020. doi:10.1001/jama.2020.20717.

[4] A. Campbell and S. Ward. Comparisons of all-cause mortality between European countries and regions: 2020. techreport, Office for National Statistics, Mar. 2021. URL `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/comparisonsofallcausemortalitybetweeneuropeancountriesandregions/2020`.

[5] Center for International Earth Science Information Network. *Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11*, 2018. Accessed 2021-01-11.

[6] Climate Prediction Center. *Global Daily Temperature*. NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, 2021. URL `https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html`. Accessed February 24, 2020.

[7] B. J. Cowling, S. T. Ali, T. W. Y. Ng, T. K. Tsang, J. C. M. Li, M. W. Fong, Q. Liao, M. Y. W. Kwan, S. L. Lee, S. S. Chiu, J. T. Wu, P. Wu, and G. M. Leung. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. *The Lancet Public Health*, 5(5):e279–e288, may 2020. doi:10.1016/S2468-2667(20)30090-6.

[8] EuroMoMo. Euromomo weekly bulletins. online, 2020. URL `https://www.euromomo.eu/`. On-going weekly reports.

[9] T. Hager. *Nager.Date – Worldwide Public Holiday*, 2021. URL `https://github.com/nager/Nager.Date`. Accessed at February 24, 2020.

[10] HMD. *Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), 2021. URL `https://mortality.org`. Accessed at February 24, 2020.

[11] H. P. i Arolas, E. Acosta, G. López-Casasnovas, A. Lo, C. Nicodemo, T. Riffe, and M. Myrskylä. Years of life lost to COVID-19 in 81 countries. *Scientific Reports*, 11(1), feb 2021. doi:10.1038/s41598-021-83040-3.

[12] D. Jdanov, V. M. Shkolnikov, A. A. Galarza, C. Boe, and M. Barbieri. Short-Term Mortality Fluctuations Dataseries methods protocol. Technical report, Max Planck Institute for Demographic Research, 2021. URL `https://mortality.org/Public/STMF_DOC/STMFNote.pdf`. Accessed March 31, 2021.

[13] V. Kontis, J. E. Bennett, T. Rashid, R. M. Parks, J. Pearson-Stuttard, M. Guillot, P. Asaria, B. Zhou, M. Battaglini, G. Corsetti, M. McKee, M. D. Cesare, C. D. Mathers, and M. Ezzati. Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries. *Nature Medicine*, oct 2020.

doi:10.1038/s41591-020-1112-0.

[14] D. A. Leon, V. M. Shkolnikov, L. Smeeth, P. Magnus, M. Pechholdová, and C. I. Jarvis. COVID-19: a need for real-time monitoring of weekly excess deaths. *The Lancet*, 395 (10234):e81, may 2020. doi:10.1016/S0140-6736(20)30933-8.

[15] C. Magnani, D. Azzolina, E. Gallo, D. Ferrante, and D. Gregori. How large was the mortality increase directly and indirectly caused by the COVID-19 epidemic? an analysis on all-causes mortality data in italy. *International Journal of Environmental Research and Public Health*, 17(10):3452, may 2020. doi:10.3390/ijerph17103452.

[16] P. Michelozzi, F. de'Donato, M. Scortichini, P. Pezzotti, M. Stafoggia, M. D. Sario, G. Costa, F. Noccioli, F. Riccardo, A. Bella, M. Demaria, P. Rossi, S. Brusaferro, G. Rezza, and M. Davoli. Temporal dynamics in total excess mortality and COVID-19 deaths in italian cities. *BMC Public Health*, 20(1), aug 2020. doi:10.1186/s12889-020-09335-8.

[17] C. Modi, V. Böhm, S. Ferraro, G. Stein, and U. Seljak. Total COVID-19 mortality in italy: Excess mortality and age dependence through time-series analysis. medRxiv, May 2020.

[18] K. Modig, A. Ahlbom, and M. Ebeling. Excess mortality from COVID-19: weekly excess death rates by age and sex for sweden and its most affected region. *European Journal of Public Health*, 31(1):17–22, nov 2020. doi:10.1093/eurpub/ckaa218.

[19] L. Németh, D. A. Jdanov, and V. M. Shkolnikov. An open-sourced, web-based application to analyze weekly excess mortality based on the Short-term Mortality Fluctuations data series. *PLOS ONE*, 16(2):e0246663, feb 2021. doi:10.1371/journal.pone.0246663.

[20] S. Rizzi and J. W. Vaupel. Short-term forecasts of expected deaths. *Proceedings of the National Academy of Sciences*, 118(15):e2025324118, mar 2021. doi:10.1073/pnas.2025324118.

[21] M. Scortichini, R. S. dos Santos, F. D. Donato, M. D. Sario, P. Michelozzi, M. Davoli, P. Masselot, F. Sera, and A. Gasparrini. Excess mortality during the COVID-19 outbreak in italy: a two-stage interrupted time-series analysis. *International Journal of Epidemiology*, 49 (6):1909–1917, oct 2020. doi:10.1093/ije/dyaa169.

[22] R. E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports*, 78(6):494, 1963. doi:10.2307/4591848.

[23] R. J. J. Soo, C. J. Chiew, S. Ma, R. Pung, and V. Lee. Decreased influenza incidence under COVID-19 control measures, Singapore. *Emerging Infectious Diseases*, 26(8):1933–1935, aug 2020. doi:10.3201/eid2608.201229.

[24] A. Stang, F. Standl, B. Kowall, B. Brune, J. Böttcher, M. Brinkmann, U. Dittmer, and K.-H. Jöckel. Excess mortality due to COVID-19 in germany. *Journal of Infection*, 81(5):797–801, nov 2020. doi:10.1016/j.jinf.2020.09.012.

[25] The Economist Data Team. Tracking COVID-19 excess deaths across countries. *The Economist*, Apr. 2020. URL `https://www.economist.com/graphic-detail/2020/04/16/tracking-covid-19-excess-deaths-across-countries`. Accessed April 16, 2020.

[26] D. M. Weinberger, J. Chen, T. Cohen, F. W. Crawford, F. Mostashari, D. Olson, V. E. Pitzer, N. G. Reich, M. Russi, L. Simonsen, A. Watkins, and C. Viboud. Estimation of excess deaths associated with the COVID-19 pandemic in the united states, march to may 2020. *JAMA Internal Medicine*, jul 2020. doi:10.1001/jamainternmed.2020.3391.

[27] S. H. Woolf, D. A. Chapman, R. T. Sabo, D. M. Weinberger, L. Hill, and D. D. H. Taylor. Excess deaths from COVID-19 and other causes, march-july 2020. *JAMA*, 324(15):1562, oct 2020. doi:10.1001/jama.2020.19545.

[28] J. Wu and A. McCann. 25,000 missing deaths: Tracking the true toll of the coronavirus crisis. *New York Times*, 2020. URL `https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html`. Accessed April 21, 2020.

Supplementary materials

Model description

i	observation within a country	α	intercept	j	sex×age stratum
t	weeks since start of data series	β, ζ	time trend coefficients	$j[i]$	stratum of observation i
w	weeks since start of epi-year	γ	seasonality coefficient	D	death count
H	public holiday indicator	δ	holiday coefficient	E	person-weeks exposure
T	temperature anomaly	ν	temperature anomaly coef.	f, g	smooth functions of time

We fit the *Serfling with exposures* (SRFr) via

$$\begin{aligned} D_i \sim & \text{Pois}(\lambda_i) \\ \log \lambda_i = & \alpha_{j[i]} + \beta_{j[i]} t_i + \\ & \gamma_{1,j[i]} \sin\left(\frac{2\pi}{52} w_i\right) + \gamma_{2,j[i]} \cos\left(\frac{2\pi}{52} w_i\right) + \gamma_{3,j[i]} \sin\left(\frac{2\pi}{26} w_i\right) + \gamma_{4,j[i]} \sin\left(\frac{2\pi}{26} w_i\right) + \\ & \delta_{j[i]} H_i + \log E_i. \end{aligned} \tag{1}$$

Omitting the term $\log E_i$ from (1) yields the *Serfling without exposures* (SRFc) model. For the *Euromomo style Serfling model* (SRFcem) we dropped the exposure, holiday and half year seasonal cycle from (1) and only fitted the model over a five year period on weeks 15 through 26 and 36 through 45.

The *Generalized Additive Model with temperature anomaly* (GAMrt) allows for non-parametric, smooth seasonal effects and adjusts for the effect of extreme temperature on weekly death counts. We specify the model as

$$\begin{aligned} D_i \sim & \text{Pois}(\lambda_i) \\ \log \lambda_i = & \alpha_{j[i]} + \beta_{j[i]} t_i + f_{j[i]}(w_i) + \nu_{w[i]} T_i + \delta_{j[i]} H_i + \log E_i, \end{aligned} \tag{2}$$

where $f_{j[i]}(w_i)$ is a stratum specific seasonality term implemented as a cyclical penalized spline, and $\nu_{w[i]} = g(w_i)$ is the temperature anomaly coefficient varying smoothly over week of year in a cyclical fashion. For the *Generalized Additive Model without temperature anomaly* (GAMr) we omitted this term.

The *Latent Gaussian model with temperature anomaly* (LGMrt) follows the specification by [13]. The model is fitted separately by age group and sex and in the following we omit the subscript $j[i]$.

$$\begin{aligned} D_i \sim & \text{Pois}(\lambda_i) \\ \log \lambda_i = & \alpha + \beta t_i + \zeta_{t[i]} + \gamma_{t[i]} + \nu_{w[i]} T_i + \delta H_i + \log E_i. \end{aligned} \tag{3}$$

The model features three components modeling the weekly time series of death rates: βt_i is a simple log-linear trend over time, $\zeta_{t[i]} \sim \text{Normal}(\phi \zeta_{t[i]-1}, \sigma_\zeta^2)$ is a first order autoregressive trend, and $\gamma_{t[i]}$ is a seasonal random effect under the prior that the summed effect of every consecutive series of 52 weeks on mortality is distributed as $\sum_{s=0}^{51} \gamma_{t[i]+s} \sim \text{Normal}(0, \sigma_\gamma^2)$, thus allowing for a varying magnitude of seasonality over years.

[To be continued]

Supplementary figures

Figure S.1: Rolling origin five-fold cross-validation setup mirroring the task of predicting weekly deaths past the beginning of the COVID pandemic given pre-pandemic data.

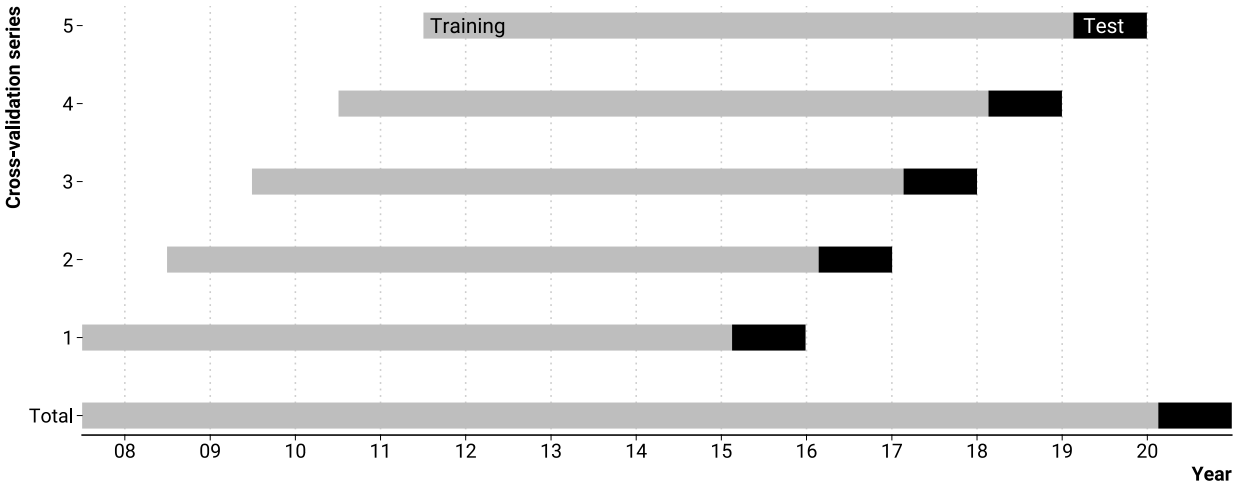


Figure S.2: Population weighted weekly temperature anomaly by country.

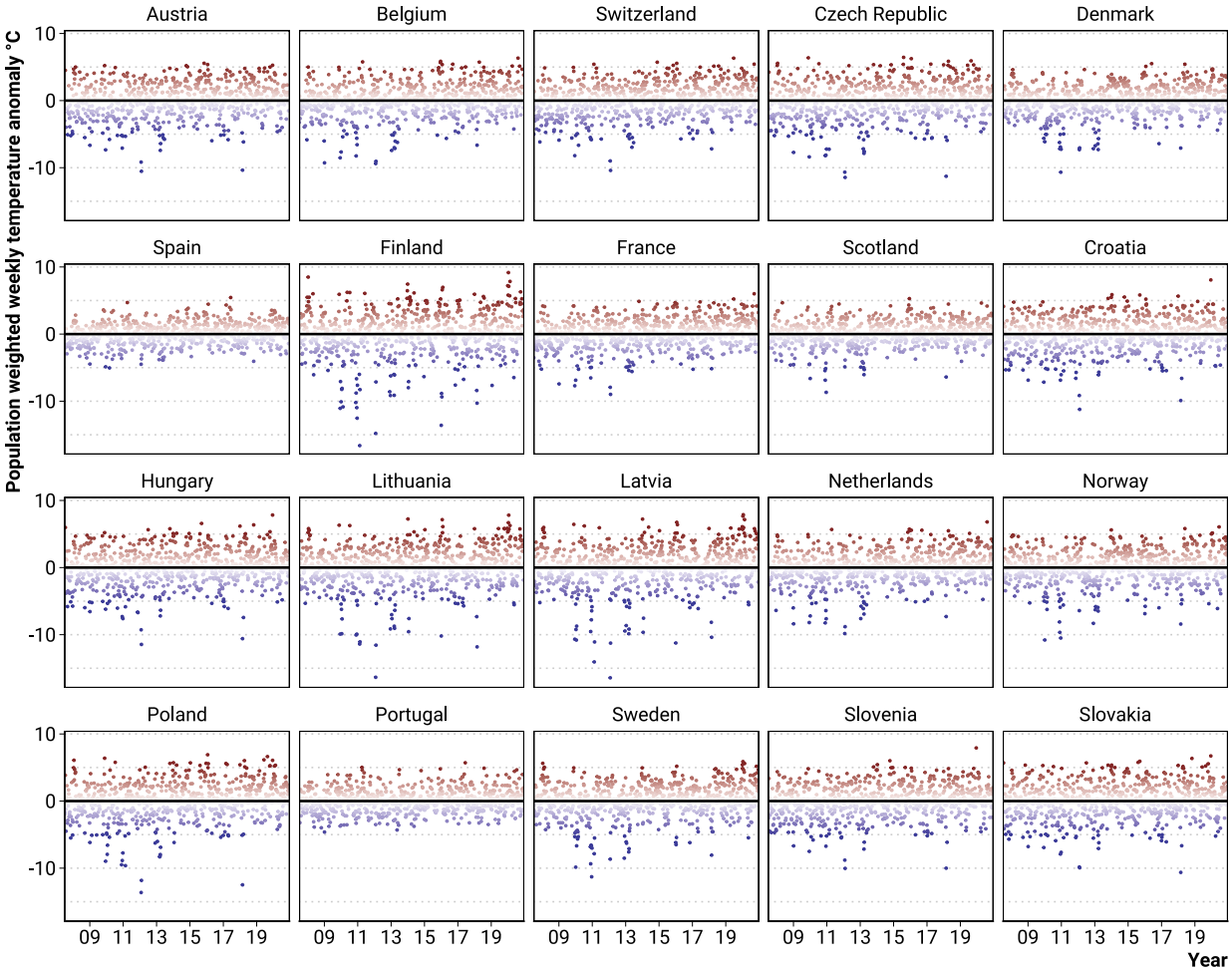


Figure S.3: Female percent excess deaths for ages 0 to 65 as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

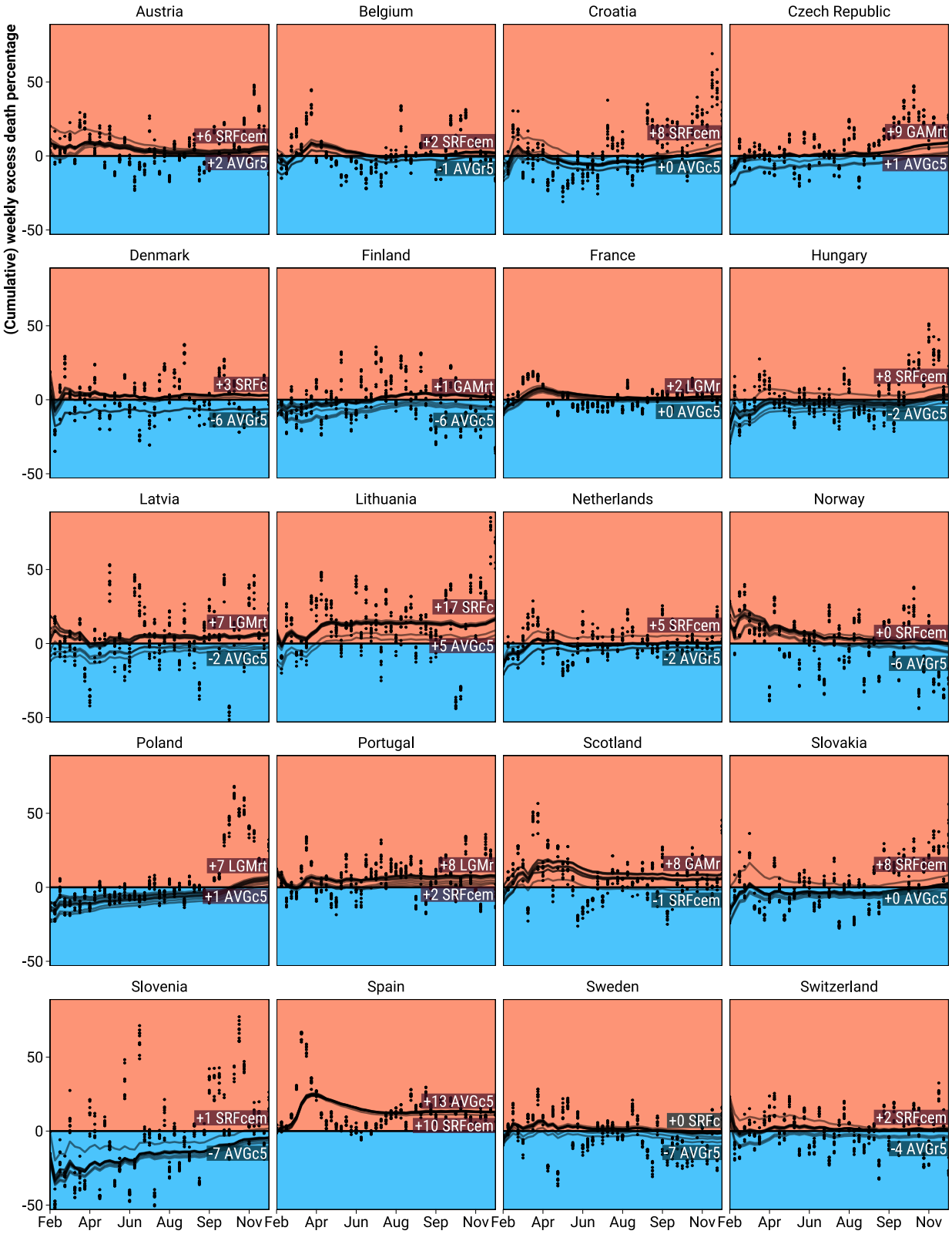


Figure S.4: Female percent excess deaths for ages 65 to 75 as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

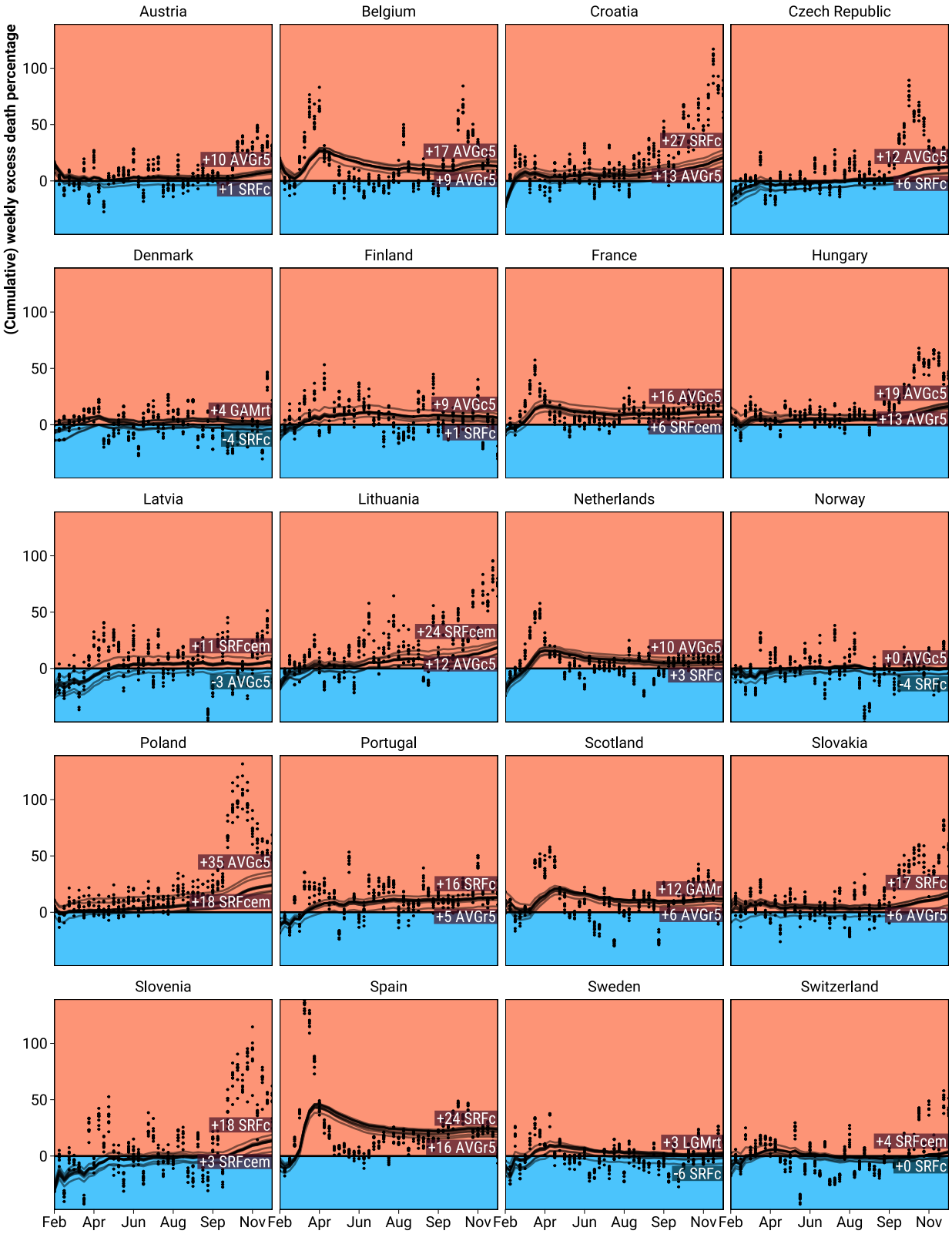


Figure S.5: Female percent excess deaths for ages 75 to 85 as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

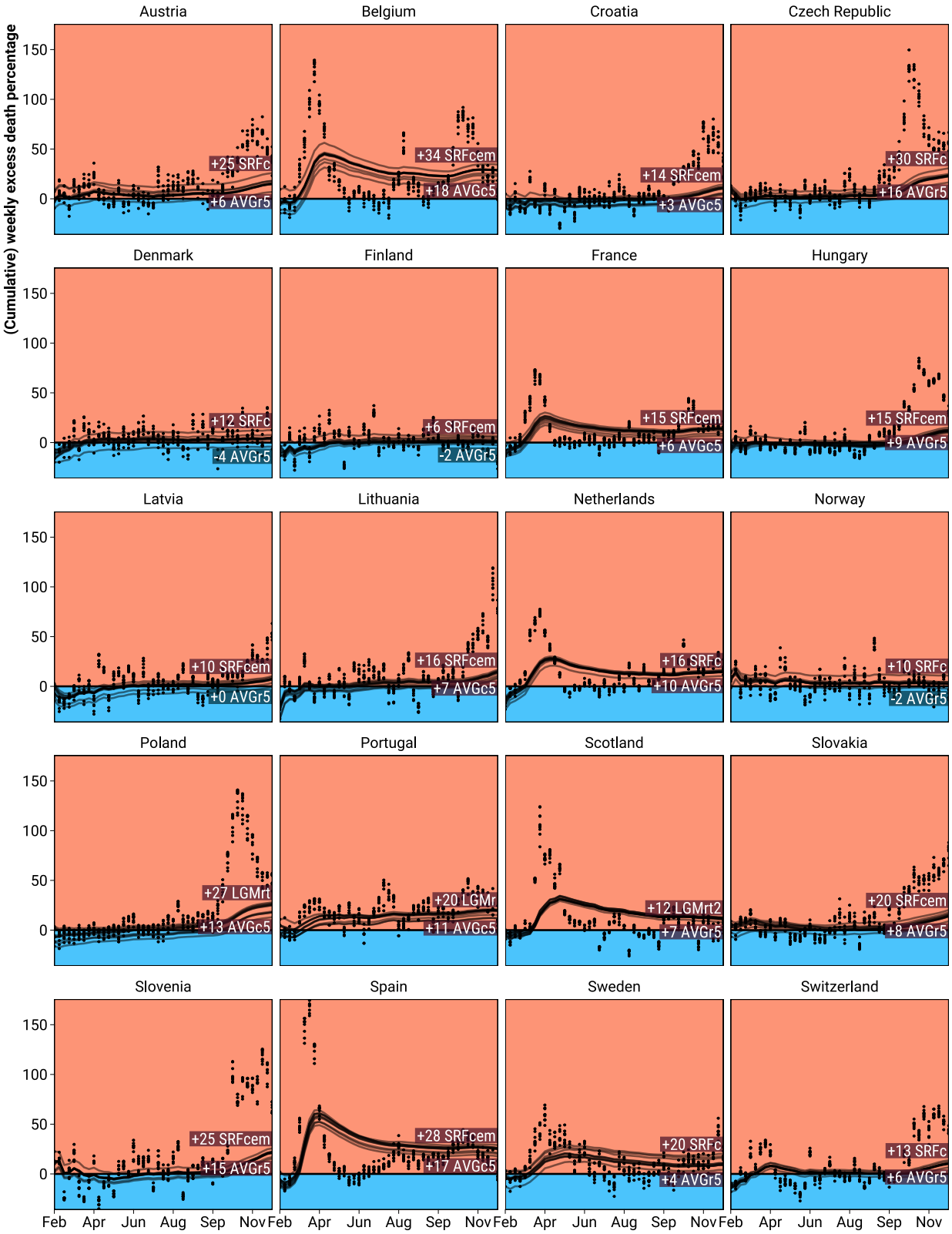


Figure S.6: Female percent excess deaths for ages 85+ as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

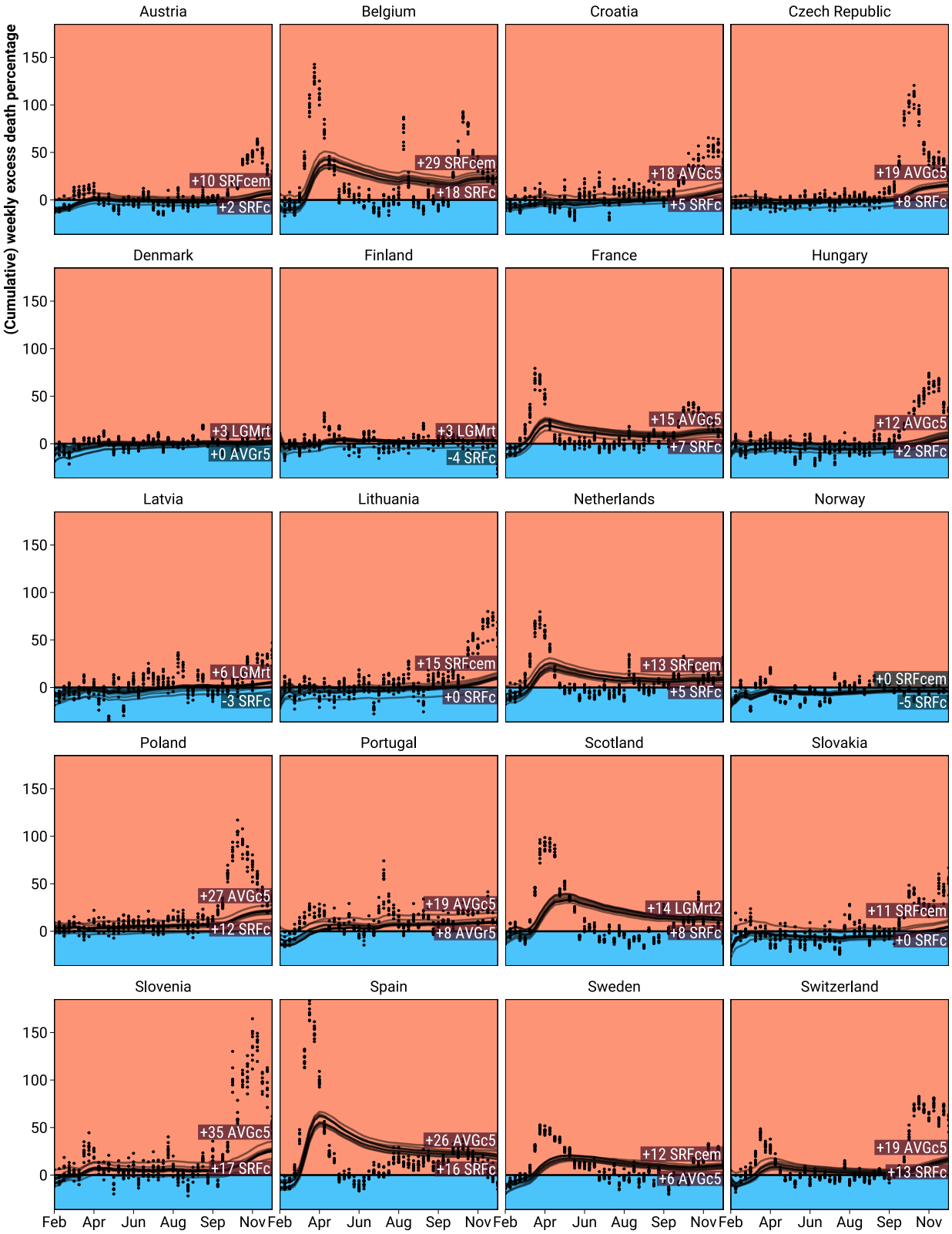


Figure S.7: Male percent excess deaths for ages 0 to 65 as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

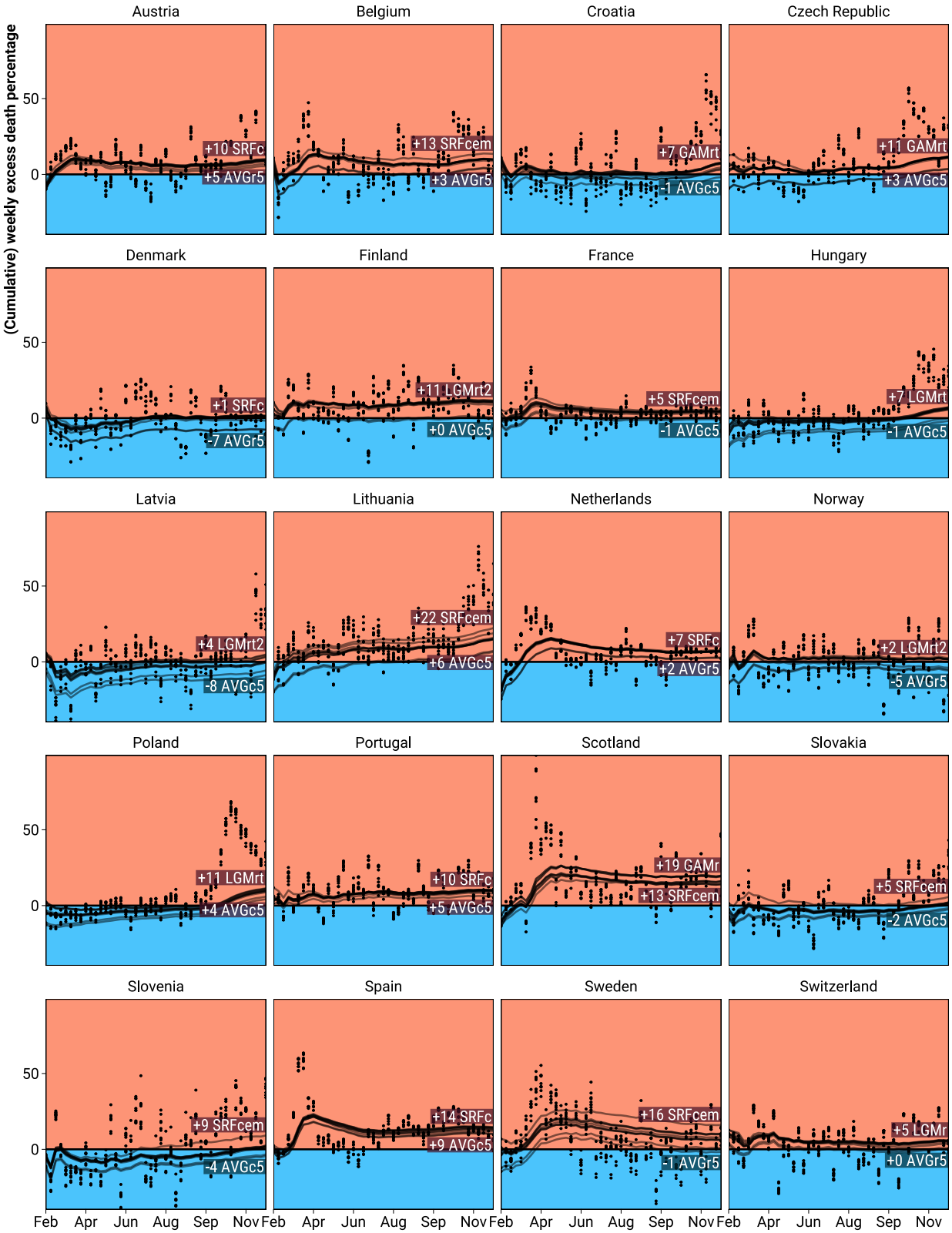


Figure S.8: Male percent excess deaths for ages 65 to 75 as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

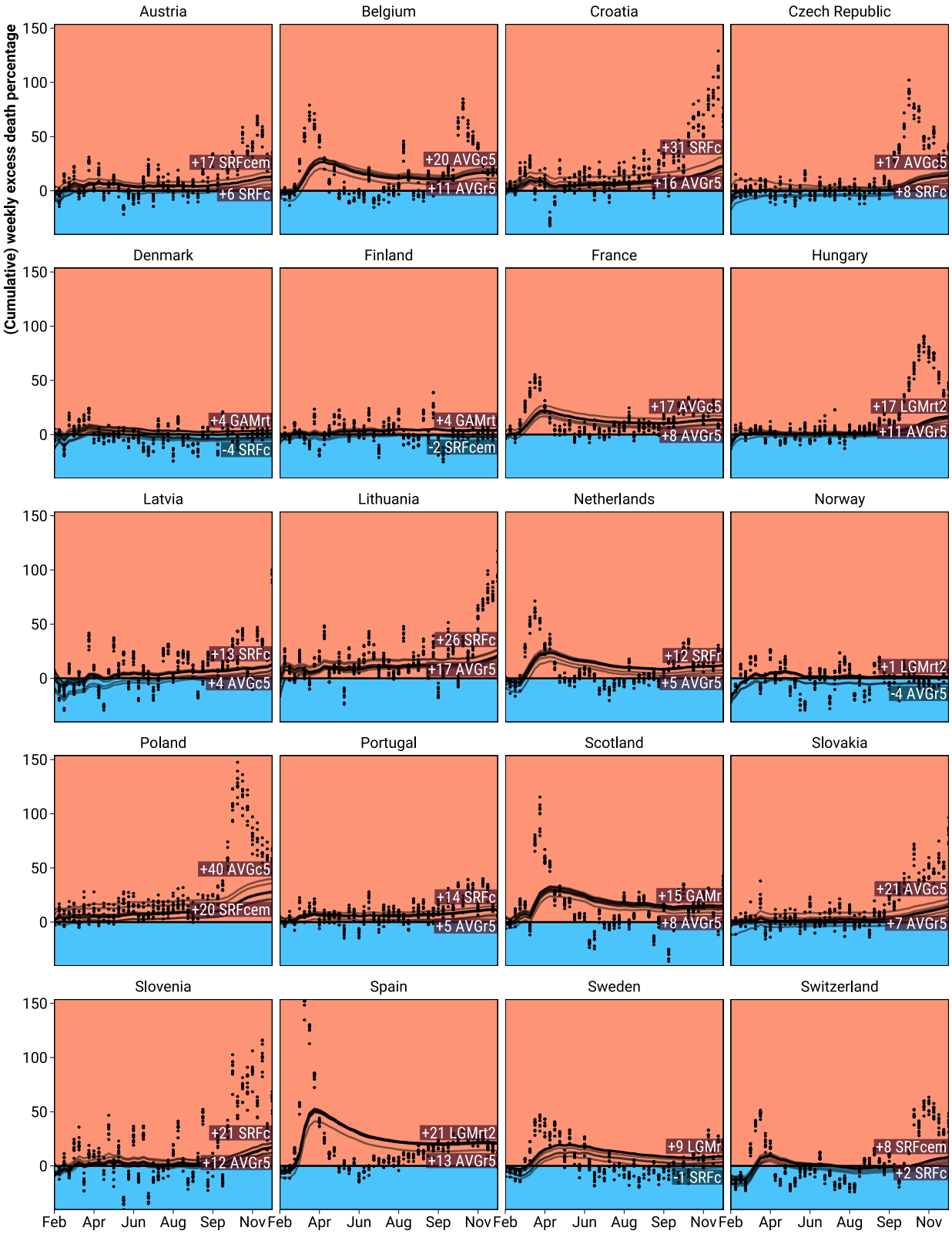


Figure S.9: Male percent excess deaths for ages 75 to 85 as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

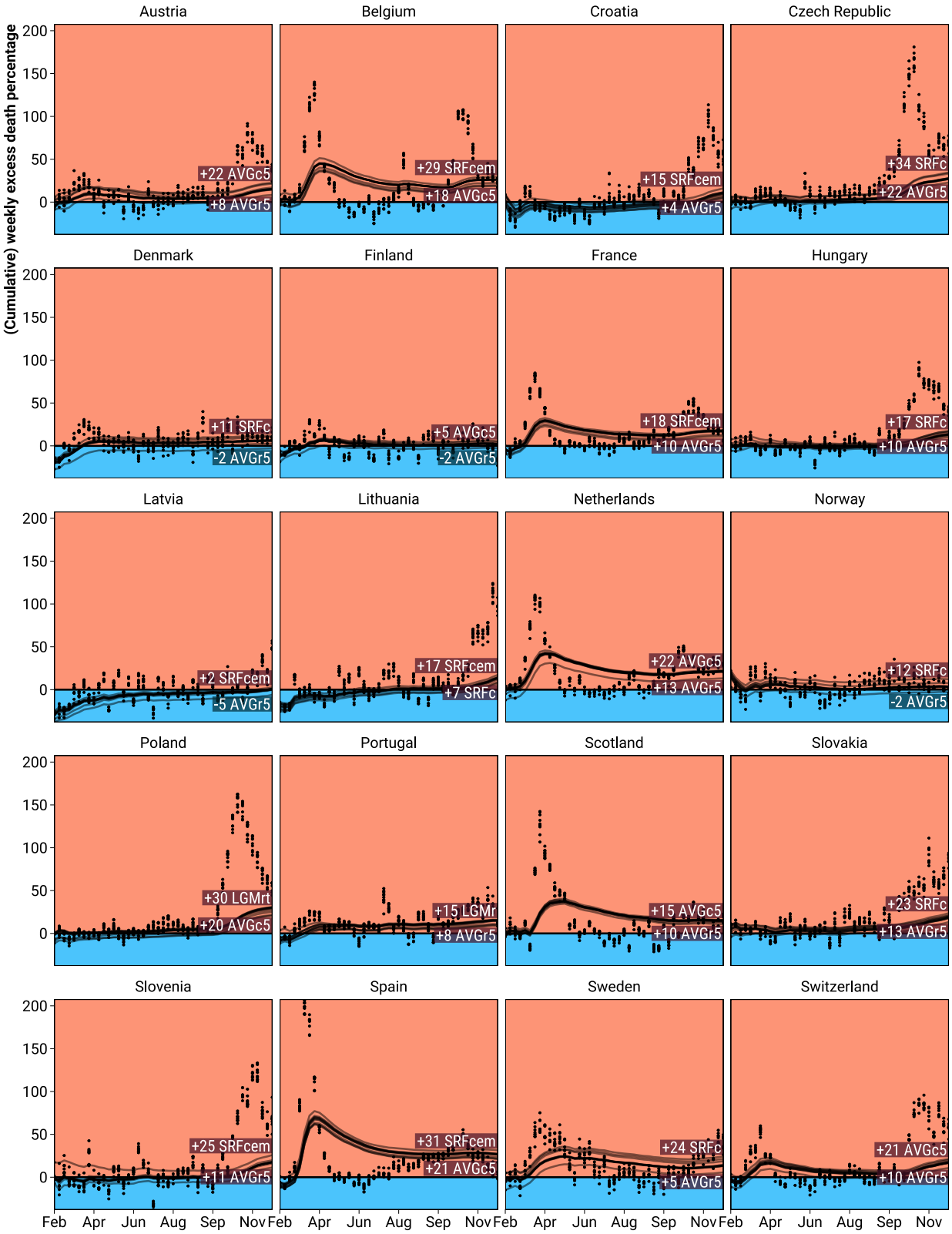


Figure S.10: Male percent excess deaths for ages 85+ as predicted from 10 different models during the year 2020 weeks 8 through 52 for 20 European regions.

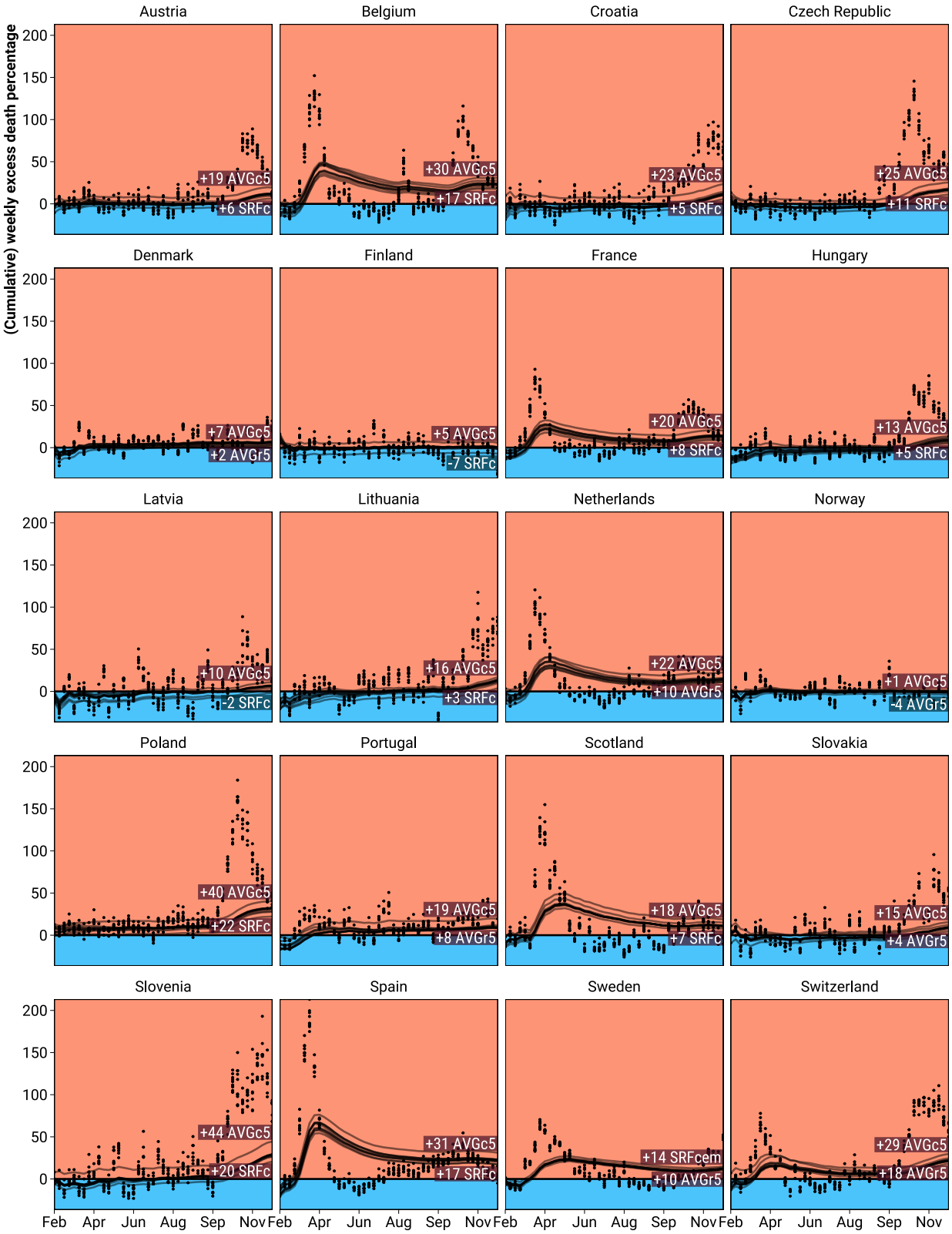


Figure S.11: Within-stratum country ranking of excess death percentage during the year 2020 weeks 8 through 52 under 10 different models.