Comparison of surveillance flu data across regions

René Ferland, Sorana Froda, Anthony Coache

ÉMoStA, Département de mathématiques, UQAM, Montréal, Québec, Canada

We consider surveillance influenza data gathered by the CDC (Center for Disease Control and Prevention, USA) collected since October 1997 until the present day. One set of reported data is weekly new cases of ILI (influenza like illness as defined by a specific protocol) reported by a surveillance network of health providers (GP and pediatric practices); the yearly number of providers varies (has increased over time). Additional information on new cases is the age group (4 or 5 classes), and HHS (Health and Human Services) regions (1-10). Given the way data are collected one cannot use the raw data to compare regions and seasons.

The main purpose of the present research is to explore a way of performing such comparisons across regions and years. Namely, rather than comparing incidence rates, we consider the relative distribution of new cases during the first 33 weeks of the flu season where the proportions p_i of new cases in week $i, i = 1, 2, \ldots, 33$ among all weeks, depend on explanatory variables. The analysis is based on the public data posted on the CDC website (*Flu View*).

Methodology

Available surveillance data (on ILINet): weekly new cases (divided in four age groups: 0-4, 5-24, 25-64, and 65+), as well as weekly visits (for any reason) and number of health providers who are reporting to the CDC; this information is given at the national and regional level, over 19 seasons (from 1997–1998 until 2016–2017) where the first week of each season is 40 (first week in October). The rates published by the CDC are adjusted to take into account the population size and we present them in Figure 1. For the first years there was no data beyond week 20 (33d week of the flu season); therefore we consider only 33 weeks in each season (October till end of May).

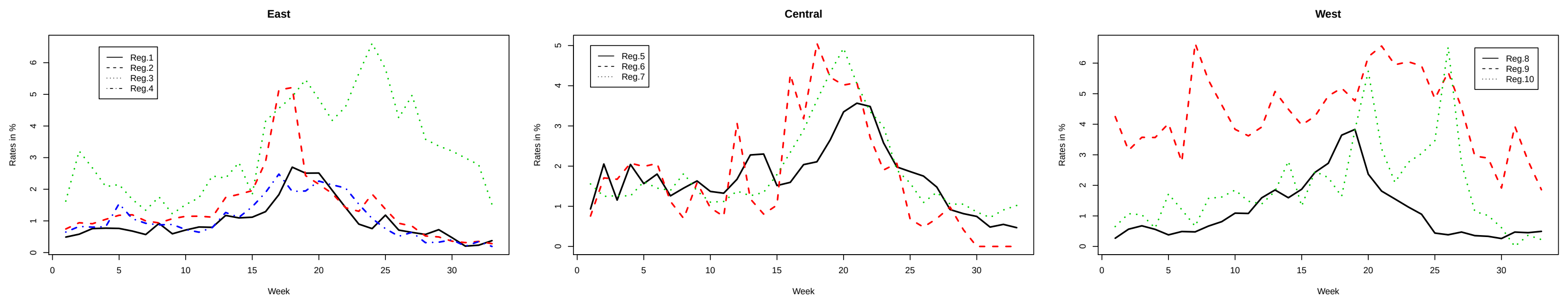

Figure 1. CDC "raw" data: weighted incidence rates (in %) in season 5 (2001-2002), grouped by: East (1–4), Central (5-7), West (8-10).

The fit: we explain the variation in incidence rates (new cases by visits) by applying a negative binomial regression model; **explanatory variables**: week (as factor), number of providers (as a proxy for yearly variation), number of cases in age group 0-4 (as a proxy for pediatric practices among the surveillance clinics); we excluded the year 2009, spring and fall (pandemic year).

The prediction: we compare the prediction given by each regional model as applied to typical generic sets of explanatory variables. For ease of comparison, we grouped HHS regions according to two criteria: population size (according to percentages of total USA population, as given by the last census) and geographical location. Over time, the number of participating health practices has increased, and we distinguished seasons by three such levels of participation.

Generic data sets: we considered the national (USA) data during three flu seasons, corresponding to different levels of providers' participation: low, season 5 (2001-2002), medium, season 10 (2006-2007), and high, season 16 (2012-2013); further, for each season, we computed the relative distribution (by week, out of 33 weeks) of the explanatory variables and created "generic" regional data; this allows to use the same generic distribution in all regions, while preserving the order of magnitude of the data in each region.

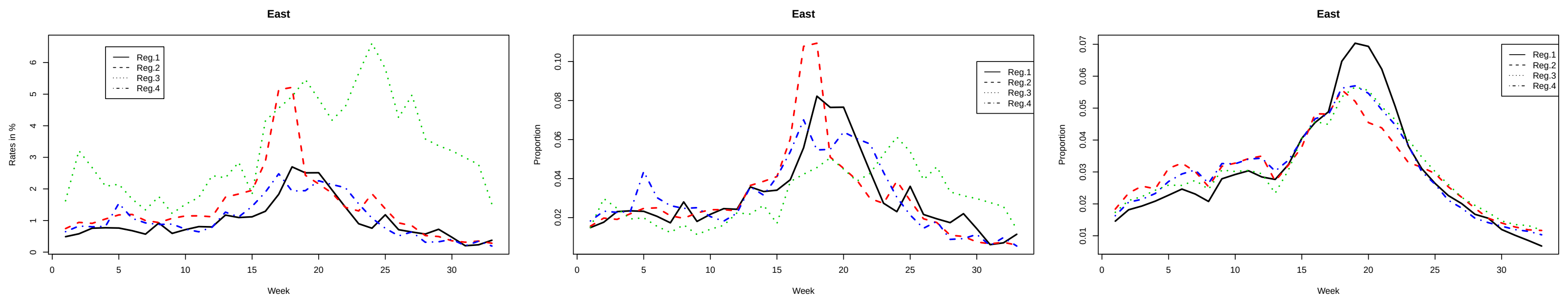

Figure 2. Three methods of comparison applied to season 5 (2001-2002), East (1–4): incidence rates (in %), relative incidence rates, and predictions (both as proportions).

Groups by population size

Sizes: Small (less than 5% of total population, 4 regions), Medium (between 9.5% and 12.8%, 3 regions), Large (more than 15.5%, 3 regions).

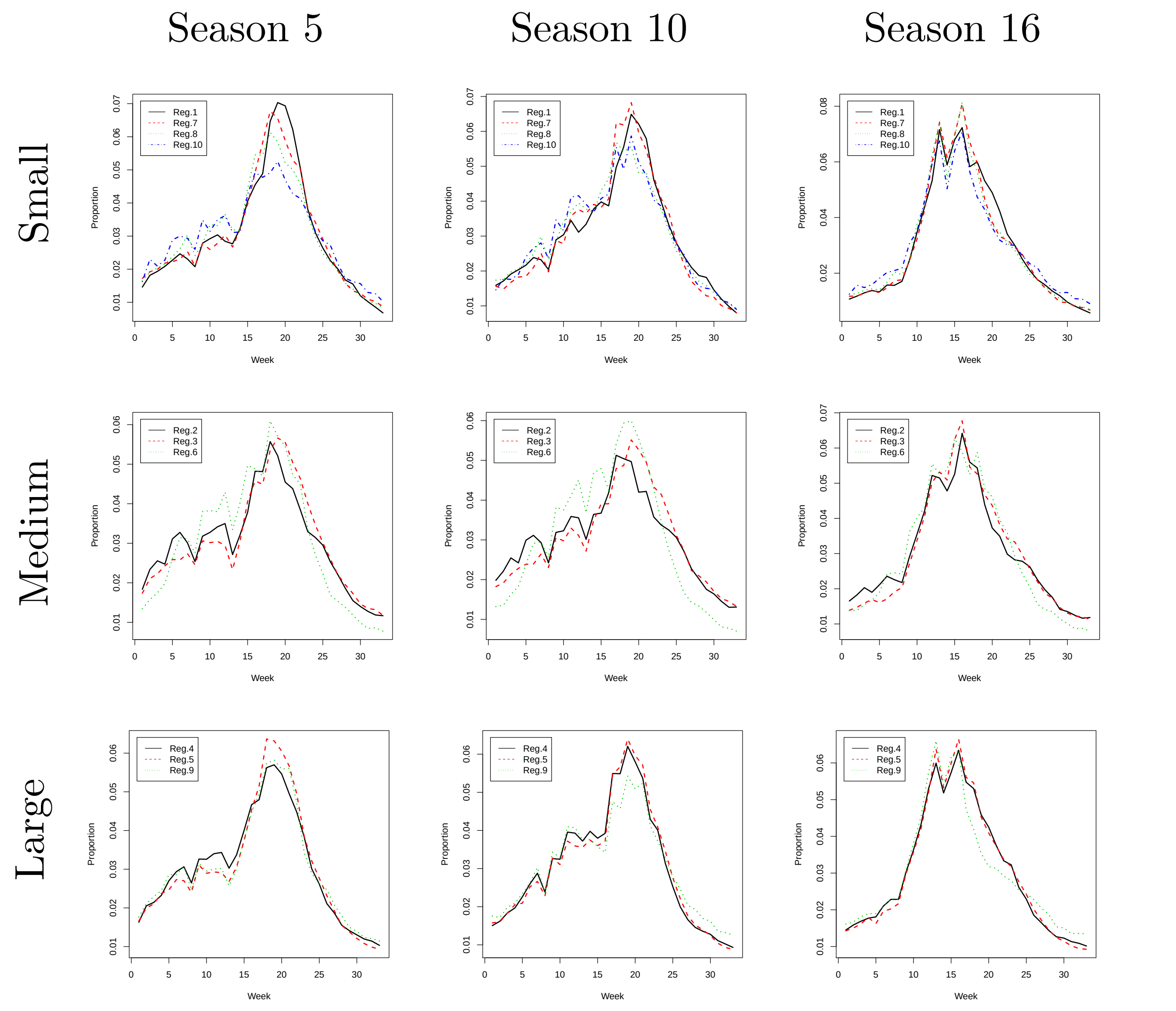

Figure 3. Predicted values for regions 1, 7, 8, and 10 (small), regions 2, 3, 6 (medium), and regions 4, 5, 9 (large) based on generic data for season 5, season 10, and season 16.

Groups by location

Locations: regions grouped into East (4 regions), Central (3 regions), West (3 regions).

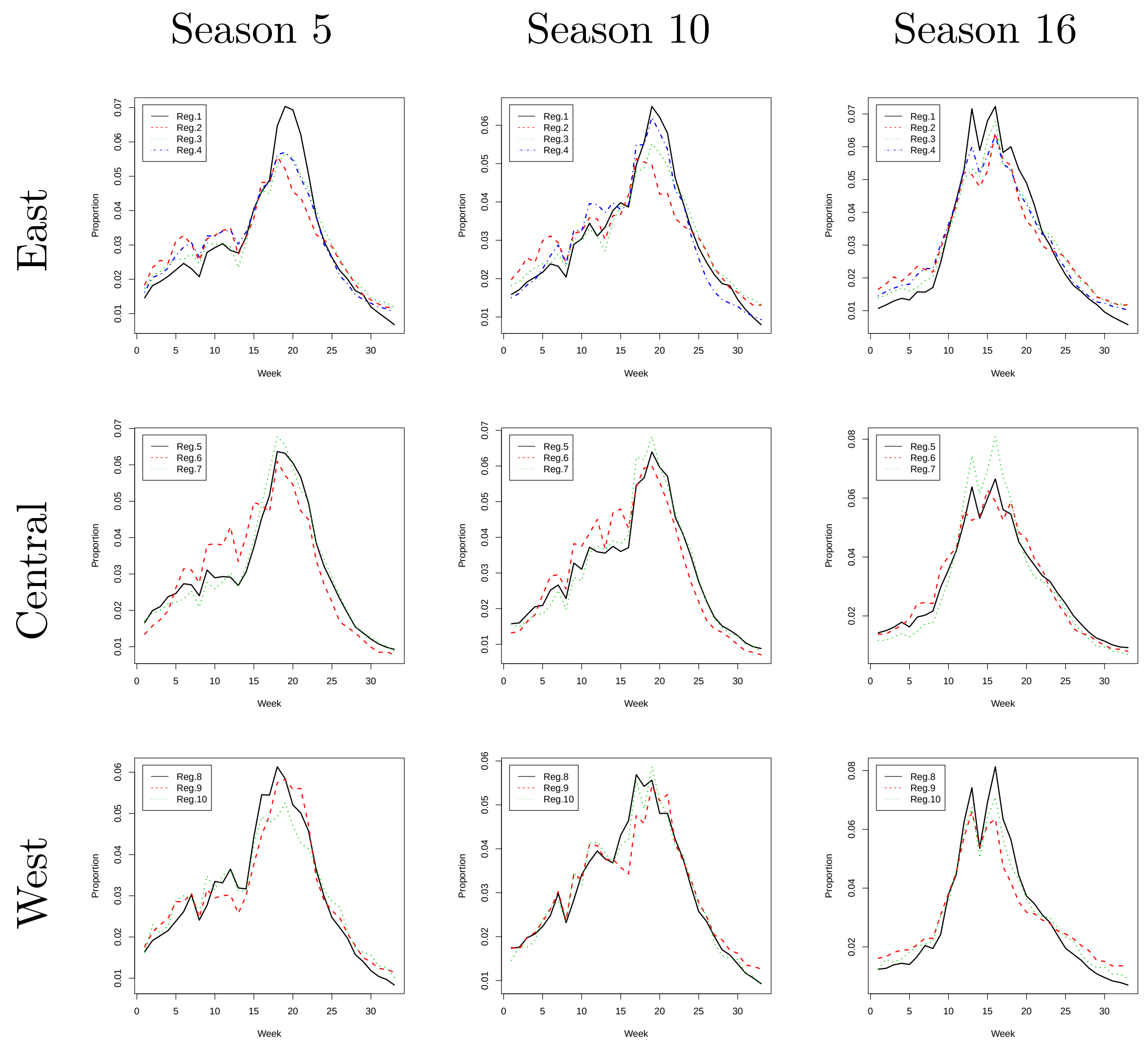

Figure 4. Predicted values for regions 1, 2, 3, 4 (East), regions 5, 6, 7 (Central), and regions 8, 9, 10 (West) based on generic data for season 5, season 10, and season 16.

Discussion

Alternative approaches for comparing flu seasons are mainly based on time series techniques applied to incident cases or mortality data (Olsen *et al.* 2013, Viboud *et al.* 2004, Viboud *et al.* 2006). Given that the seasonal component is an accepted fact, **our idea** is to consider each season separately and try to take advantage of the available data on additional variables. As for the progression over the years, we make use of explanatory variables in order to take it into account.

Generic sets: definitely other choices are possible, for example one could use partially reconstructed data from a multivariate analysis of the regions; a useful application is to consider data in an "unusual year", like a pandemic year, in order to asses its differences with regular flu seasons. This type of question is of great interest to epidemiologists and public health practitioners (see, e.g., Chuang *et al.*).

Future developments: Huppert *et al.* have performed comparisons by resorting to time series techniques and fitting a special SIR model to spatially homogeneous data. On the other hand, Froda and Leduc (1914) have developed a stochastic SIR model based on (time) non homogeneous Poisson and birth-death processes in order to estimate R_0 (the basic reproduction number). This last model could be extended in a natural way in order to incorporate explanatory variables.

As far as the **data analysis** goes, we would like to take into account other information that is available on the CDC website, like the distribution of virus strains in a given year. Moreover, hospitalizations, laboratory tests, or mortality data in 122 main cities (due to influenza and related conditions) could be treated in a similar way.

Bibliography (selected papers). ⊡ Froda, S. and Leduc, H., 2014. Estimating the basic reproduction number from surveillance data on past epidemics. *Math. Biosci.* 256, 89–101. ⊡ Chuang, J. H., Huang, A. S., Huang, W. T., Liu, M. T., Chou J. H., Chang, F. Y., and Chiu W. T., 2012. Nationwide Surveillance of Influenza during the Pandemic (2009-10) and Post-Pandemic (2010-11) Periods in Taiwan. *PLoS-One 7(4).* ⊡ Huppert A., Barnea O., Katriel G., Yaari R., Roll U., and Stone L., 2012. Modeling and Statistical Analysis of the Spatio-Temporal Patterns of Seasonal Influenza in Israel. *PLoS ONE 7(10).* ⊡ Olsen, D.R., Konty, K.J., Paladini, M., Viboud, C., and Simonsen, L., 2013. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Comput Biol 9(10).* ⊡ Viboud, C., Boëlle, P.-Y., Pakdaman, K., Carrat, F., Valleron, A.-J., and Flahault, A. 2004. Influenza Epidemics in the United States, France, and Australia, 1972-1997. *Emerging Infect. Dis.* 10, 858—875. ⊡ Viboud C., Bjørnstad O. N., Smith D. L., Simonsen L., Miller M., and Grenfell B. T., 2006. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science.* 312, 447-451.

Congrès annuel de la SSC ⋆ Annual Meeting of the SSC ⋆ 11 Juin -14 June, 2017 ⋆ Winnipeg