CSE-8803 Project Final Report: Leveraging Regional Similarity in Deep Epidemic Forecasting

Alexander Rodriguez*
Georgia Institute of Technology
Atlanta, GA, USA
arodriguezc@gatech.edu

Rogelio Rodriguez*
Georgia Institute of Technology
Atlanta, GA, USA
rrodriguez77@gatech.edu

ABSTRACT

Forecasting emerging pandemics such as COVID-19 using purely data-driven models is challenging, in no small part because data is very sparse. In such a scenario, purely data-driven models need to leverage as much relevant data is available. For example, leveraging similarity between regions based on geographical closeness can be helpful but we should not limit ourselves to this type of relations. Therefore, we propose to leverage similarity between regions based on a latent space build from 'static' and 'dynamic' features from each region. Static features for each region include demographics, social behavior, hospital capacity, comorbidities, and socioeconomic data. On the other hand, dynamic features are signals such as mobility, hospitalizations, confirmed cases, etc. By projecting regions into a temporal latent space (i.e. for each region and time frame, we have a different latent representation), we can leverage both static and dynamic similarity. Our specific goals are as follows. (1) Find in the literature epidemiologically relevant static and dynamic information; (2) develop a neural architecture to leverage static and dynamic similarity. In our experiments, we found that incorporating region similarity can help our forecasting accuracy in some cases, especially for 1-week ahead.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

Deep learning

ACM Reference Format:
Alexander Rodriguez and Rogelio Rodriguez[1]. 2019. CSE-8803 Project Final Report: Leveraging Regional Similarity in Deep Epidemic Forecasting. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Motivation. It has been approximately 9 months since the first atypical pneumonia cases in Wuhan city, Hubei province, China [20]. SARS-CoV-2, the pathogen causing COVID-19 disease has infected 54,256,914 and killed 1,315,369 people [7] as of November 15th, 2020. Without a highly effective treatment nor a vaccine to prevent the disease, the world faces difficult times. This crisis is not only a global health crisis but also a social and economic one, which has wreaked havoc on the world in both developed and developing countries.

Accurately predicting key indicators of the spread of infectious disease provide valuable information to plan mitigation efforts; however, it remains a major challenge. There are many complexities to take in account, such as social behavior, weather, and interventions. In addition to that, the still chaotic scenario, uncertain social landscape, and the scarce available data have potential to mislead any model. In the epidemic forecasting literature, mechanistic models [14, 19] are built on assumptions on how the disease spreads. On the other hand, statistical models [2, 6] rely on recognition of patterns and correlations. While successful in several forecasting tasks in influenza,COVID-19 poses a greater challenge as the data for training these models is even more scarce than for other diseases. Therefore, in a way to mitigate data scarcity, we propose to explicitly model two facets of regional similarity: dynamic (time series signals) and static (stationary characteristics of a region).

Currently, there are several statistical models that leverage static features such as demographic and socioeconomic factors to assess potential risks associated with COVID-19 in a specific region. However, the majority of the statistical models surveyed for this proposal [9, 13, 16, 18] do not consider relevant regional epidemiological information. Only one [9] uses regional epidemiological data despite the importance of this information for assessing risk for worse COVID-19 evolution at the patient level [3, 4]. Hence, we see this gap of incorporating epidemiological and socially relevant information into statistical models as an opportunity to leverage this relevant information at the regional level. We propose to do so by learning similarity among regions to improve predictions of our deep epidemic forecasting model.

Approach and Contributions. Our main idea is that learning and using regional similarities should help in both forecasting and interpretation. Our neural architecture will learn how to embed static and dynamic similarity among regions. Learned regional similarity will help in the forecasting of different regions. To our best knowledge, this would be the first work in this regional similarity for forecasting.

2 RELATED WORK

Epidemic Forecasting. Modeling approaches for epidemic forecasting can be broadly categorized in mechanistic [14] and statistical [2, 6]. In the recent years, statistical models have been the most successful in several forecasting targets [**?**]. In influenza forecasting, various statistical approaches have been proposed: Bayesian modeling [5], Kernel methods [6], Gaussian processes [21], and ensembles of mechanistic and statistical methods [10]. More recently, the deep learning community has take interest in this problem in

*Equal contribution.

Conference'17, July 2017, Washington, DC, USA
2019. ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

forecasting influenza [2, 11, 17, 21] and COVID-19 [9, 12]. Deep learning approaches have an advantage of being capable of ingesting data from multiple sources and being able to tell what the data says without many assumptions [12]. Ramchandani et al. [9] have explored the use of static features; however, this nor previous work have focused on explicitly modeling regional similarity.

Epidemiological, Demographic, and Socioeconomic Factors. There is still an unknown number of factors that affect the COVID-19 epidemic outcome in a given region. However, statistical models have started to use demographic, epidemiological, and socioeconomic information in an attempt to understand/identify the factors matching with regional variation in COVID-19 risk. Sannigrahi et al. [13] found that the uneven distribution of COVID-19 cases and deaths across 31 European countries can be explained by differences in sociodemographic factors between the countries. They also found strong associations between cases and country income levels and between deaths and the country's total population. Whittle and Diaz-Artiles [18] found that variations in COVID-19 test positivity rates among neighborhoods in New York City can be explained by sociodemographic factors such as dependent youth population, population density, income and race/ethnicity community proportions. Verhagen et al. [16] forecast the expected number of hospitalizations at the county level for England and Wales using demographic data and pre-COVID hospital bed capacity data. They were able to identify areas at increased risk of facing health care burdens due to COVID-19 by including socioeconomic factors into their statistical model. Finally, Ramachandani et al. [9] developed a deep learning model to forecast the range of increase in COVID-19 cases in a given day at the county level (for all US counties). In their model, they use dynamic and static heterogeneous data such as census data, intra and inter county mobility, social distancing, cases, and epidemiological data. From all statistical models surveyed for this work, only the model of Ramachandani et al. [9] uses region-specific epidemiological data, setting a precedent for our work that will use similar information. In stark contrast with case/clinical studies showing that epidemiological factors are important for assessing the potential risk associated with COVID-19 [3, 4], the majority of statistical models have failed to use this type of information.

3 PROBLEM FORMULATION

Forecasting Targets. The forecasting target is reported incidence (new) deaths for US states and the US overall as reported by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) [7],which serves as gold standard for the CDC.

Problem Formulation. We can state our forecasting problem as follows. **Given**: for each region $r \in \mathcal{R}$, we have a static feature vector $\mathbf{s}^r$, an observed multivariate time series of COVID-related signals $\mathcal{X}^r = \{\mathbf{x}_i^r\}_{i=1}^N$, corresponding values for the forecasting target $\mathcal{Y}^r = \{y_i^r\}_{i=1}^N$, where N is the size of the sequence until the current date, and a set of tuples of given similarities $\mathcal{G}$. **Predict:** next k values of forecasting target, i.e. $\{\hat{y}_i^r\}_{i=N+1}^{N+k}$.

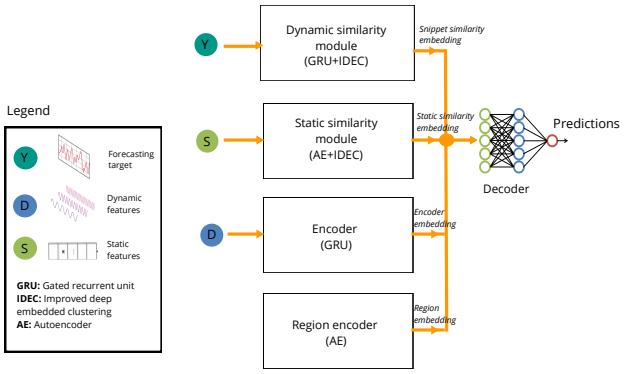

Figure 1: Schematic of our proposed end-to-end framework

4 METHODS

In our method, we will leverage similarity from static and dynamic features. In doing so, our neural architecture (depicted in Fig. 1) will contain a snippet segmentation module, dynamic similarity module, static similarity module, encoder, and decoder.

Dynamic Similarity Module. Improved Deep Embedded Clustering (IDEC) [8] has been proved to be a successful unsupervised method for learning low-dimensional representations with a similarity objective function. Previous successful applications of this method have been used in EpiDeep, a recent deep forecasting model for influenza [2] based on seasonal similarity. This method leverages similarity by mapping an embedding of a snippet to the full seasonal sequence. We describe this method in more detail in our appendix.

In this module, we encode the sequence via a GRU (gated recurrent unit) and then apply the deep clustering to the latent representations of each sequence.

Static Similarity Module. Here, we can again leverage IDEC in a more straightforward manner as the static feature vector is a fixed-length vector. In this case, we do not need to do a mapping as we do for the dynamic similarity module. To deal with the set of given similarities $\mathcal{G}$, we will add a loss that encourages to place two regions close in the learned latent space.

Region Encoder. We will have an autoencoder to obtain region embeddings from 1-hot encodings of each region.

Encoder. The encoder is a recurrent neural network that takes a multivariate sequence of signals and learns long- and short-term dependencies. The output of this module will be an embedding, which will be concatenated with the embeddings from the similarity modules.

Decoder. In this module we leverage embeddings from similarity modules, region embedding, and encoder to make predictions. For this, we use a feedforward network that will output the k-th prediction.

5 EXPERIMENTS

5.1 Dataset collection

We use two datasets, classified by the temporal nature of their features: dynamic

Dynamic Regional Data: For dynamical features, we will use the COVID-related signals described in detail in [12], which comes from multiple sources and have been categorized as line-list-based, testing-based, crowdsourced symptomatic, mobility, exposure, and social surveys. *Data format:* For each region, we have a weekly time series (for dynamic data) and a regional feature vector (static data). *Dataset size:* Each region contains weekly records starting in April 2020. Thus, we have approximately a sequence of 30 observations per signal per region.

Static Regional Data: We collect our regional static features from different sources. For example, from US Census Bureau[1] we collected demographics (e.g., median age, population, population density) and socioeconomic (e.g., poverty rate, income inequality) information. We obtained the number of adults per state ($\geq$ 18 yrs old) with $\geq$ 1 comorbidities that increase the risk for COVID-19 disease complication from [1]; additional factors exacerbating risk for COVID disease complication were obtained from the GitHub repository of the Yu group at UC-Berkeley [2]. Social behavior data including results from mask wearing surveys were obtained from the New York Times repository[3]. Additional social behavior data such as fraction of people spending +6h away from home per day was obtained from the COVIDcast site[4] and data on average daily trips and fraction of people staying at home was obtained from the United States Department of Transportation[5]. *Data size and format:* We have 24 static features for 51 US states.

5.2 Static Regional Data Analysis

The main idea of this project is to learn and use static and dynamic regional similarity to improve forecasting and interpretation of model predictions. Hence, for the static module we performed an initial exploratory analysis to learn similarities between US states based on the static information available for each state. For the analysis, we included features that we thought were relevant in assessing potential COVID risk at the state level. We grouped our set of static variables into five main groups: demographics, comorbidities, hospital capacity, social behavior, and socioeconomic (see Table 1 for the comprehensive list of static features in each group). We started by exploring the correlations between the cumulative number of deaths and positive COVID cases as of October 24th, 2020, and demographic and comorbidity features. Results show (Fig. 2) that the total number of COVID cases and deaths are highly correlated with the state population (PoP, $r \geq 0.84$) and slightly correlated with population density (PoPDe, $r \geq 0.13$). On the other hand, variables within the comorbidities category correlate with each other ($r \geq 0.53$) but not with the number of COVID cases and deaths. Comorbidities variables included in this analysis are: number of adults with chronic conditions (Adchr), diabetes percentage (Diabe), heart disease (Heart), and chronic respiratory disease (RespM) mortality. It is interesting the negative correlation between chronic respiratory disease mortality and the number of COVID deaths and cases.

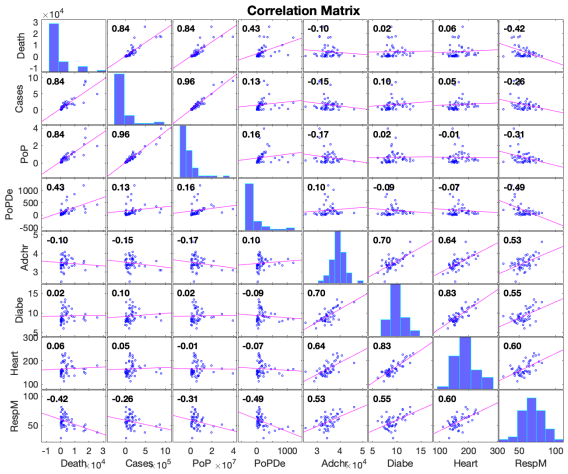

Figure 2: Correlation Matrix: Static features vs cumulative COVID cases and deaths

To identify how similar states are between each other based on the static features describing them, we performed a k-means clustering analysis. States within each cluster are similar among them and dissimilar when compared to states in a different cluster. The clustering analysis was performed (1) using the complete set of static features (24 features) or (2) using a subset of the dataset containing features only for the corresponding category (see Table 1 for the different categories). Before performing the analysis we standardized our data set, $(\frac{x-\mu}{\sigma})$, so that each variable contributed equally to the analysis. We identified $k = 3$ as the optimal number of clusters producing a smaller average within-cluster sum of squares. We further performed a principal component analysis (PCA) in order to visualize our clusters in 2D.

Results of the clustering analysis projected onto the first two components of the PCA are summarized Fig.3. The three clusters distinguish from each other when projected in the PCA. Cluster 1 is the largest with 25 states while clusters 2 and 3 both have 13 states each. To further help with the visualization, we project the clusters obtained by k-means onto the US map so we can observe clearly what states are within each cluster. It is interesting, the geographical pattern recapitulated by cluster 2 (south-east states) and 3 (mid-west and part of north-west states), while for cluster 1 regions span the north-east, south-west, and the west.

To analyze how specific categories of static regional data impact the clustering outcome we performed a k-means clustering analysis using subsets rather than the complete data set. The variables used for this analysis correspond to the static categories listed in Table 1. The results of the analysis are summarized in Fig. 4, showing the clusters for the different categories projected on the US map. The social behavior data category produces clusters dividing the country almost perfectly by west coast (magenta), center & south (green), and east coast (blue). By taking a look at the data, states in the green cluster tend to have greater mobility with a smaller percentage of people staying at home, a larger average number of daily trips, and

[1] https://www.census.gov/programs-surveys/acs/data.html
[2] https://github.com/Yu-Group/covid19-severity-prediction
[3] https://github.com/nytimes/covid-19-data/tree/master/mask-use
[4] https://covidcast.cmu.edu/
[5] https://www.bts.gov/daily-travel

Table 1: Static variables

Categories	Variables	Rationale
Demographics	1. Total population; 2. Population density;3. Population in the 65+ age group; 4. Median age	Elderly people are more likely to have severe infections and more likely to require hospitalizations [4]. Additionally, greater population density has been proposed to be linked to greater transmission rates of SARS-CoV-2 associated, likely due to increasedcontact rates [15, 18]
Comorbidities	5. Number of adults with ≥ 1 chronic conditions increasing risk of COVID complication; 6. Percentage of Diabetes; 7. Heart disease mortality rate; 8. Stroke mortality rate; 9. Percentage of adult smokers;10. Chronic respiratory disease mortality rate.	CDC study [1] using information from China, showing that adults with chronic conditions such as cardiovascular disease, diabetes, chronic respiratory disease among others have higher case-fatality rate.
Hospital capacity	11. Number of hospitals; 12. Number of ICU beds; 13. Number of doctors	Evidence of hospital capacity per state works as an instrument to measure state's preparation for COVID-19 crisis
Social behavior	14. Proportion of people who respond 'Never' to mask usage survey; 15. who respond 'Rarely'; 16. who respond 'Sometimes'; 17. who respond 'Frequently'; 18. who respond 'Always' 19. Percent of people staying at home; 20. Average daily trips 21. Fraction people spending +6h away from home	Evidence of implementation of non-pharmaceutical and usefulinterventions such as mask usage at the state level
Socioeconomics	22. Social Vulnerability Index (SVI) percentile ranking 23. Poverty rate; 24. Income inequality	Measuring the impact of poverty, lack of access to transportation, crowded housing, etc. on community preparedness to respond to hazardous events.

a larger fraction of people staying 6 or more hours away from home compared to sates in the other two clusters. When socioeconomic factors are used for the clustering analysis, the US map is divided into two, south (magenta) and north (blue and green). It interesting, that states in the magenta cluster tend to have greater levels of income inequality compared to states in the other two clusters. The demographics category produces less clear geographical clusters but it is clear that some of the most populous states are clustered together (green cluster: NY, TX, CA). On the other hand, the hospital capacity data category recapitulates a similar clustering pattern as to when using the whole data set (compare Fig. 4 top-right and bottom-right). Until we have direct evidence that a specific category matches with regional COVID outcomes, we will continue using our complete static data set to inform our forecasting model.

5.3 Forecasting Results

5.3.1 Setup. All experiments are conducted with a Linux machine of 40 processors Intel Xeon CPU E5-2698 v4 @ 2.20GHz, with 252 GB of RAM. We coded the deep learning architecture in Pytorch. Training of the complete architecture with data from all regions for a single future target takes around 3 minutes. All the results are based predictions during three months (July to October 2020, epidemic weeks 30 to 42).

5.3.2 Research Questions. We address these two research questions in our forecasting experiments.

Q1. Our primary research question is to understand whether or not incorporating region similarity as in our approach leads to better results. Therefore, we want to compare against a models that do not explicitly incorporate this information.

Q2. DeepCOVID [12] is a deep learning method that ingests the same dynamic signals as ours. It is trained for a single region, for which it trains multiple weak learners (bootstrap) and average them to make point predictions. Can we beat DeepCOVID with this new neural architecture?

5.3.3 Evaluation and Baseline. For measuring performance, we use the mean absolute error, i.e., MAE $= \frac{1}{N}\sum_{i=1}^{N} \mid e_i \mid$ and its value describes how large on average the error is.

To understand the impact of incorporating region similarity, we include as a baseline a subset of our model into the evaluation. This contains only the encoder, regional embedding, and decoder. As

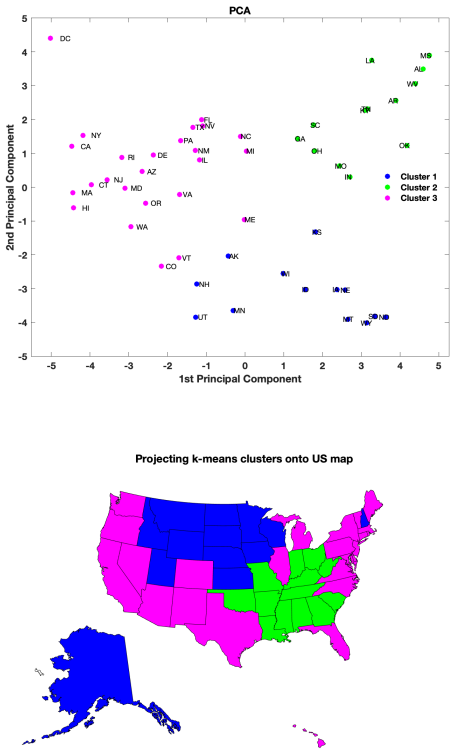

Figure 3: Top, PCA and $k = 3$ clustering analysis. Bottom, projection of cluster on the US map.

mentioned before, we also compare against DeepCOVID. We utilize its publicly available predictions found at the COVID-19 Forecast Hub.

5.3.4 Observations and Findings. We summarize our results in Table 2, in which we bold the cells where by incorporating similarity our performance improves. We noted that similarity was more effective for 1-week ahead, while for 2-week ahead is usually not

Table 2: Forecasting results for US National, California, Georgia, Texas and Pennsylvania

Model	Target	US	CA	GA	TX	PA
DeepCOVID	1wk	770.14	120.07	71.98	227.37	34.82
	2wk	942.34	143.68	81.70	241.43	43.08
Encoder + regional embedding	1wk	1755.66	193.74	109.86	370.73	210.80
	2wk	786.49	156.12	90.66	261.64	104.48
Our Method	1wk	**1050.11**	**129.84**	**73.88**	**340.23**	279.19
	2wk	1038.09	**156.11**	113.28	341.32	144.60

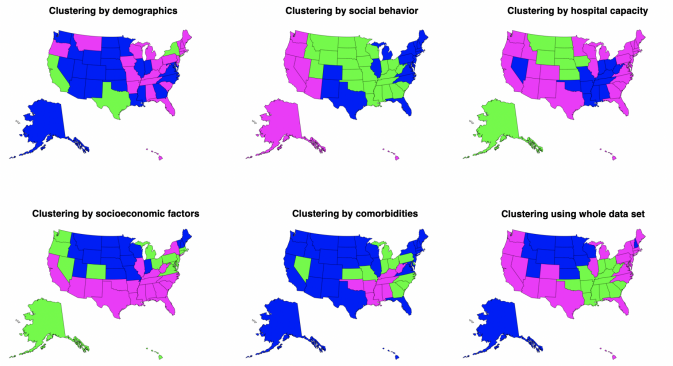

Figure 4: Results of clustering states using various categories of static regional data. Map at the bottom right is the result of using the whole dataset.

helping. This is very interesting and we find it hard to explain, but our intuition says that it is most likely artifact of our dynamic similarity module, which exploits the most recent trends in data across all regions. Further investigation is needed to fully understand this effect of regional similarity.

We also found that our proposed architecture is unable to beat DeepCOVID. This may be due to the fact that DeepCOVID's weak learners are more robust than a single large architecture. For the future, we could explore how to use bootstrap with our architecture so that ensemble it to have better point estimates and also enable uncertainty estimation.

6 DISCUSSION

In this work, we presented a novel method to incorporate regional similarity to improve epidemic forecasting. This is an important piece of information because data is very scarce in this domain, thus, we want to take advantage of similar patterns that happened in other similar regions. Our results indicate that this is a promising source of data, but we need to do more experiments to fully understand how similarity representations help our forecasts. We also did extensive clustering analysis that can be expanded to clusters learned by our deep clustering modules. The regional static data analysis allowed us to identify how US states cluster together based on 24 state-level static features. Static variables cover a diverse set of categories, including socioeconomic, demographic, social behavior, comorbidities, and hospital capacity factors. The recapitulation of geographical clusters by our k-means clustering analysis is exciting (Fig. 3). Results can be interpreted in several ways, e.g., (1) geographical closeness can be relevant in assessing potential COVID risk, especially for states in clusters 1 and 2, (2) a regional rather than a state-level response might be more adequate, i.e., allocating a similar amount of resources among states in cluster 1 and among states in cluster 2 given similar necessities in those regions. Moreover, from looking at the projected clusters on the map (Fig. 3-bottom) we could relate to the temporality of the COVID pandemic in the US. Some of the states in cluster 1 (e.g., WA, CA, NY) were hit early by the pandemic in spring, while states in cluster 2 (e.g., GA, AL, TN) were hit during the summer and now in the fall states in cluster 3 (e.g., SD, ND, WI) are seeing an explosion in COVID-19 cases. We are not implying that COVID temporality in the US is caused by state-level differences in some static features rather our analysis encourages further research to identify if the static regional differences have an actual impact on the outcome of the epidemic at the state level.

REFERENCES

[1] Mary L Adams, David L Katz, and Joseph Grandpre. 2020. Population-based estimates of chronic conditions affecting risk for complications from coronavirus disease, United States. *Emerging infectious diseases* 26, 8 (2020), 1831.
[2] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B. Aditya Prakash. [n. d.]. EpiDeep: Exploiting Embeddings for Epidemic Forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19* (2019). ACM Press, 577–586. https://doi.org/10.1145/3292500.3330917
[3] Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, et al. 2020. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nature Communications* 11, 1 (2020), 1–9.
[4] Ashish Bhargava, Elisa Akagi Fukushima, Miriam Levine, Wei Zhao, Farah Tanveer, Susanna M Szpunar, and Louis Saravolatz. 2020. Predictors for Severe COVID-19 Infection. *Clinical Infectious Diseases* (2020).
[5] Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. [n. d.]. Flexible Modeling of Epidemics with an Empirical Bayes Framework. 11, 8 ([n. d.]), e1004382. https://doi.org/10.1371/journal.pcbi.1004382
[6] Logan C. Brooks, David C. Farrow, Sangwon Hyun, Ryan J. Tibshirani, and Roni Rosenfeld. [n. d.]. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. 14, 6 ([n. d.]), e1006134. https://doi.org/10.1371/journal.pcbi.1006134
[7] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20, 5 (2020), 533–534.
[8] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. [n. d.]. Improved Deep Embedded Clustering with Local Structure Preservation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (2017-08). International Joint Conferences on Artificial Intelligence Organization, 1753–1759. https://doi.org/10.24963/ijcai.2017/243
[9] Ankit Ramchandani, Chao Fan, and Ali Mostafavi. [n. d.]. DeepCOVIDNet: An Interpretable Deep Learning Model for Predictive Surveillance of COVID-19 Using Heterogeneous Features and Their Interactions. 8 ([n. d.]), 159915–159930. https://doi.org/10.1109/ACCESS.2020.3019989
[10] Evan L. Ray, Krzysztof Sakrejda, Stephen A. Lauer, Michael A. Johansson, and Nicholas G. Reich. [n. d.]. Infectious disease prediction with kernel conditional density estimation: Infectious disease prediction with kernel conditional density

estimation. 36, 30 ([n. d.]), 4908–4929. https://doi.org/10.1002/sim.7488
[11] Alexander Rodriguez, Nikhil Muralidhar, Bijaya Adhikari, Anika Tabassum, Naren Ramakrishnan, and B Aditya Prakash. [n. d.]. Steering a Historical Disease Forecasting Model Under a Pandemic: Case of Flu and COVID-19. ([n. d.]).
[12] Alexander Rodriguez, Anika Tabassum, Jiaming Cui, Jiajia Xie, Javen Ho, Pulak Agarwal, Bijaya Adhikari, and B Aditya Prakash. [n. d.]. DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting. ([n. d.]). Publisher: Cold Spring Harbor Laboratory Press.
[13] Srikanta Sannigrahi, Francesco Pilla, Bidroha Basu, Arunima Sarkar Basu, and Anna Molter. 2020. Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach. *Sustainable cities and society* 62 (2020), 102418.
[14] J. Shaman and A. Karspeck. [n. d.]. Forecasting seasonal outbreaks of influenza. 109, 50 ([n. d.]), 20425–20430. https://doi.org/10.1073/pnas.1208772109
[15] Karla Therese L Sy, Laura F White, and Brooke E Nichols. 2020. Population density and basic reproductive number of COVID-19 across United States counties. *MedRxiv* (2020).
[16] Mark D Verhagen, David M Brazel, Jennifer Beam Dowd, Ilya Kashnitsky, and Melinda C Mills. 2020. Forecasting spatial, socioeconomic and demographic variation in COVID-19 health care demand in England and Wales. *BMC medicine* 18, 1 (2020), 1–11.
[17] Lijing Wang, Jiangzhuo Chen, and Madhav Marathe. [n. d.]. DEFSI: Deep Learning Based Epidemic Forecasting with Synthetic Information. 33 ([n. d.]), 9607–9612. https://doi.org/10.1609/aaai.v33i01.33019607
[18] Richard Stuart Whittle and Ana Diaz-Artiles. 2020. An ecological study of socioeconomic predictors in detection of COVID-19 cases across neighborhoods in New York City. *medRxiv* (2020).
[19] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. [n. d.]. Forecasting Seasonal Influenza Fusing Digital Indicators and a Mechanistic Disease Model. In *Proceedings of the 26th International Conference on World Wide Web - WWW '17* (2017). ACM Press, 311–319. https://doi.org/10.1145/3038912.3052678
[20] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* (2020).
[21] Christoph Zimmer and Reza Yaesoubi. [n. d.]. Influenza Forecasting Framework based on Gaussian Processes. In *Proceedings of the 37th International Conference on Machine Learning* (2020). 9.

APPENDIX

6.1 IDEC Details

As we said, for deep clustering we use IDEC. Let q_{ij} be the similarity between the embedding $\mathbf{z}_i$ and cluster center μ_j (center for cluster j).

$$q_{ij} = \frac{\left(1 + ||\mathbf{z}_i^t - \mu_j||^2\right)^{-1}}{\sum_j \left(1 + ||\mathbf{z}_i^t - \mu_j||^2\right)^{-1}}$$

where target distribution p_{ij} is

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j \left(q_{ij}^2 / \sum_i q_{ij}\right)}$$

Then, we want to minimize the distance between these two distributions via the KL divergence. Therefore, our objective in IDEC is

$$L_c^t = KL(P||Q) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

This loss will be added to our other losses (reconstruction and prediction).

6.2 Implementation Details

Data processing. It required us to scale our dynamic data within region because the data from one region is in a very different scale from others, even if the dynamics are similar. To input it to the recurrent neural network, we constructed a fixed-size tensor of the variable-length prefixes and used masking to obtain the right value of the hidden representation.

For our static data, we described in Section 5.2 how we converted our data into a scale where values across regions are comparable. However, they were still generating high losses that could impede a good optimization of other smaller losses, therefore, we scaled all regions together.

Hyperparameter tuning. For all the GRUs (encoder and dynamic similarity module), we used embeddings of size 128 to represent hidden states. For the embeddings resulting from the autoencoders, we used vectors of size 32. The decoder has four linear layers with batch normalization and dropout in the first two layers. We used leakly ReLU in the first three layers and ReLU in the last layer (output) to always obtain positive predictions. Finally, we found that having a learning rate scheduling worked best for our optimization, which goes from $1e^{-3}$ to $1e^{-5}$.