Predicting Daily Increase in COVID-19 Cases Using Dynamic Clustering With Polynomial Regression

Alexander Davidson
CSCE 4143, Data Mining
Fall 2020 Semester
University of Arkansas
Fayetteville, AR, USA
alexander.davidson98@gmail.com

***Abstract*—The novel coronavirus disease 2019 (COVID-19) has spread rapidly throughout the world since its first reported case in 2019. As of December 2020, over 70 million cases and 1.5 million deaths have been reported [3]. As a result, the pandemic has sparked a new, widespread need for knowledge discovery and prediction using the available data. In this project, I applied Dynamic Clustering with Polynomial Regression (DyCPR) to a dataset of daily increases in Coronavirus cases. The results show that for some US states, DyCPR performs better than Moving Average in predicting the number of new COVID-19 cases on the day following a limited subsequence of the time-series data.**

***Index Terms*—DyCPR, evolving clustering algorithm, polynomial regression, chaotic time-series data, coronavirus, data mining, knowledge discovery**

I. Introduction

The goal of this project was to implement a model for predicting future values in time-series data on COVID-19. The chosen datasets describe daily increases in COVID-19 cases for US states, and DyCPR was used to predict the increase in cases for a particular state on day $t+1$ given the data up to day t. DyCPR, originally introduced by Widiputra, et al., has been shown to predict future time-series values with higher accuracy than other methods such as Moving Average, and it is particularly effective on chaotic data, such as daily stock price changes [4]. Some US states, such as New York, have exhibited fairly stable growth in daily increases in cases (Fig. 1). Other states, such as Texas, display chaotic changes over small and large periods of time (Fig. 2). Because of the chaotic nature of these data, DyCPR seemed to be an appropriate model for prediction.

II. Data Preprocessing

Two datasets were used in this project:

- The primary dataset, obtained from The COVID Tracking Project, tracks daily COVID-19 statistics for 50 US states (and Washington D.C.) in 2020 [2]. Over 40 attributes are present in the dataset, but the only relevant attributes are $date$, $state$, and $positiveIncrease$ (number of new positive COVID-19 tests recorded on $date$ in $state$). The available data span from January 22 to December 11, but only dates from July 01 onward were used in

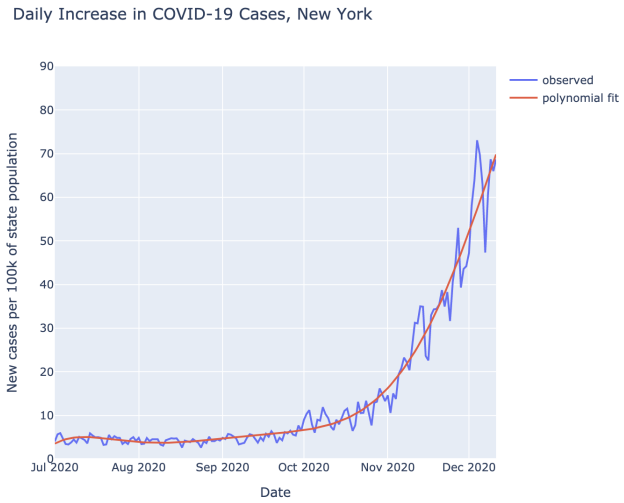

Fig. 1. New York data exhibits relatively stable growth in the daily increase in COVID-19 cases.

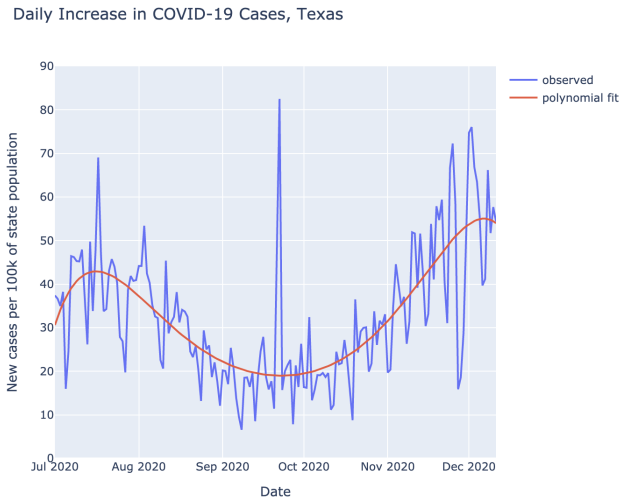

Fig. 2. Texas data exhibits relatively chaotic changes in the daily increase in COVID-19 cases.

this study due to inconsistencies in reporting and limited amounts of patient test kits in the early stages of the pandemic.

- Another dataset, obtained from the United States Census Bureau, was used to transform the former data in terms of state population [1]. The $positiveIncrease$ attribute was transformed to reflect the number of new cases recorded per 100,000 of state population over 18 years old[1]. This transformation establishes a benchmark for comparing results of DyCPR on COVID-19 data from different US states.

The datasets were inserted as tables into an SQL database, and a script was used to perform the transformation mentioned above, producing a dataset with attributes $date$, $state$, $state\_id$, $positive\_increase\_per\_100k$.

Even when considering only the data recorded after July 01, a small fraction of rows from the COVID-19 dataset were missing $positiveIncrease$. In the $positive\_increase\_per\_100k$ attribute of the final dataset, these NULL values were populated with the value from previous date for that state.

III. Model Implementation

A brief description of DyCPR is given below (adapted from the original paper proposing the algorithm) [4]. Python was used to implement DyCPR in this study:

- **Step 1:** Initialize the samples:
  The input data $\{x_1, \ldots, x_N\} \in X$ are presented as a stream of continuous values.
  - Choose $n << N$. n will be the length of subsequences (henceforth referred to as "slices") of the time-series data.
  - Produce $N - n$ slices X_i. For example:
    $$X_1 = \{x_1, ..., x_n\}, X_2 = \{x_2, ..., x_{n+1}\}, \ldots$$
    Likewise, produce corresponding slices Y_i of size $n + 1$.
  - Produce sets PF, PFM using polynomial regression on the X_i, Y_i, respectively. For each slice, determine best polynomial fit up to a chosen degree k using $minBIC_k$ (minimum Bayesian Inference Criterion)[2].
- **Step 2:** Repeat until all PF_i have been processed:
  - If no clusters exist in C, set $Cc_j = PF_i$ (centroid of C_j), $PFM_j = PFM_i$ ("best-fit regression function for next movement" of C_j). PFM_j will evolve as a superposition of similar PFM_i.
  - Find Cc_j with the minimum cosine distance to PF_i. If this distance is greater than 2 times a predetermined threshold $Dthr$, create a new cluster with centroid PF_i and best-fit regression function for next movement PFM_i, and will be used in the predictive step.
  - For each cluster C_j, set PFM_j equal to a superposition of PF_i belonging to the cluster.
- **Step 3:** To predict value x_{t+1} from a test sample $X_t = \{x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-(n-1)}\}$ of size n, find the best-fit polynomial regression function PF_t from X_t. Find the most similar cluster C_j using cosine distance between Cc_j and PF_t. Then, replace the first coefficient (y-intercept) of PFM_j with 0, and evaluate $py_{t+1} = PFM_j(n + 1)$. Then the predicted value is $x_{t+1} = x_t + py_{t+1}$.

IV. Experimentation

The data from each US state were used to train and test 10 instances of the model, independent from other states. The attributes $positive\_increase\_per\_100k$ were used for the data stream, and for each instance of the model, $N = 163$ $n = 10$, $Dthr = 0.05$, and the maximum polynomial regression degree $k = 6$. Also in each instance, 80% of the data stream slices were randomly selected for training, and the remaining 20% were used for testing.

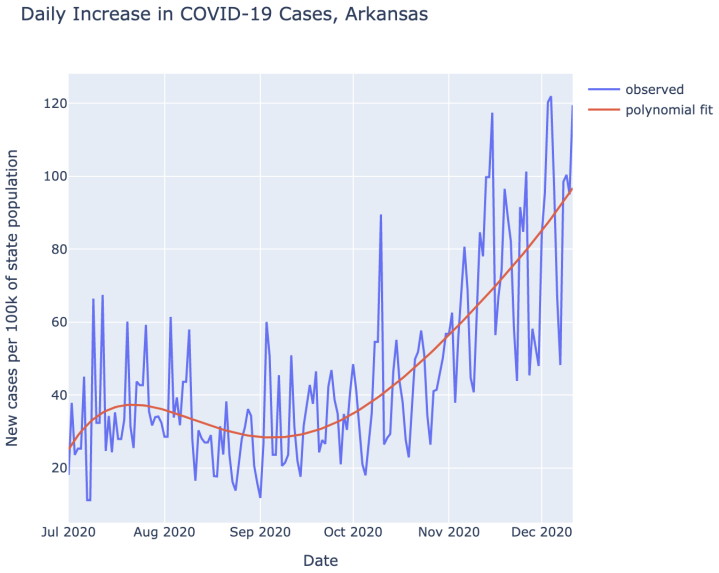

Fig. 3. Time-series data for daily increase in COVID-19 the cases in Arkansas.

[1]As of December 2020, the most recent population estimates available from the US Census Bureau were published in July 2019.

[2]The original method uses minimum mean standard error to determine best fit, but the algorithm produced more accurate predictions with $minBIC_k$ in this study

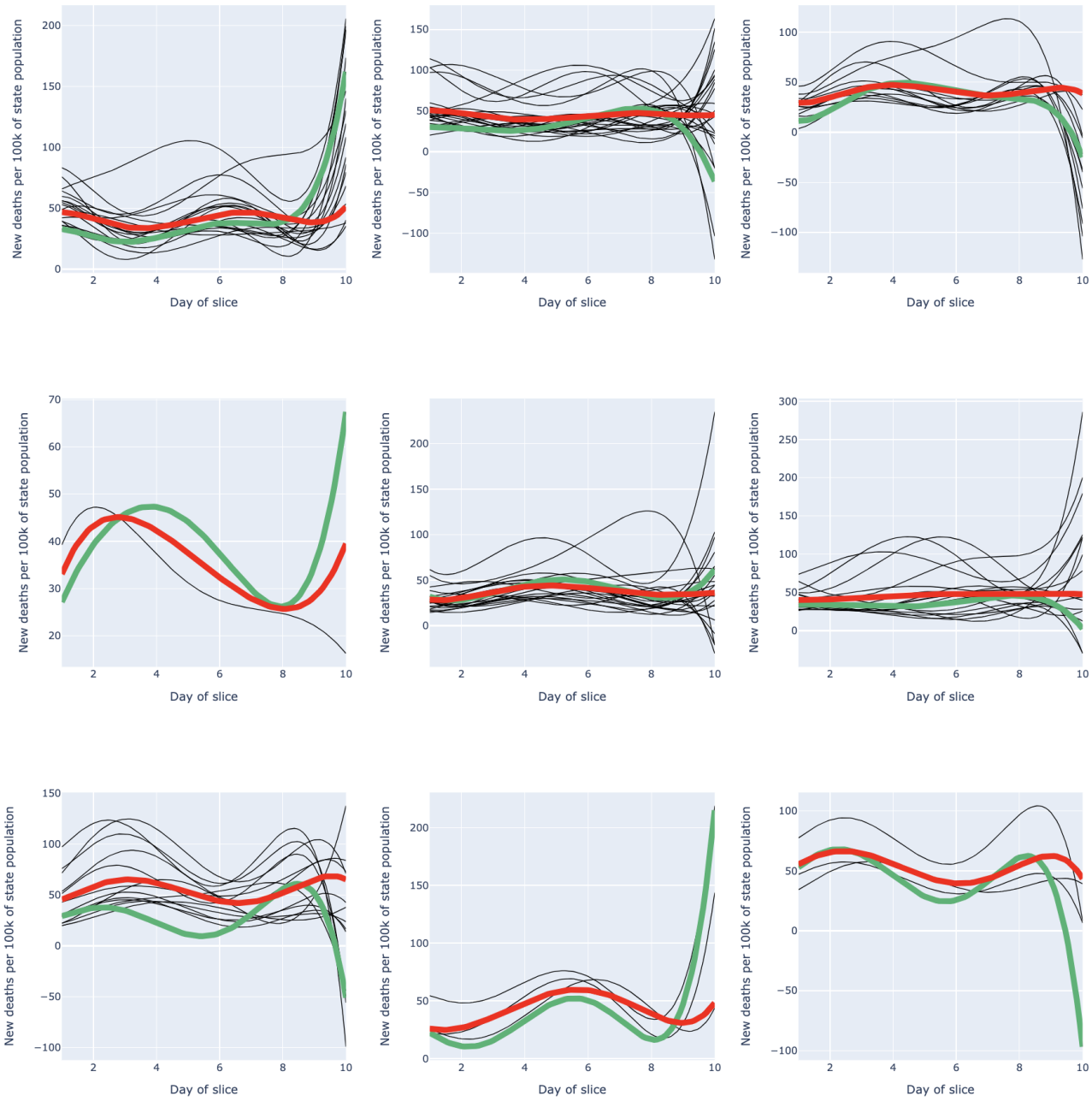

Fig. 4. Clusters produced by polynomial regression in one training instance, using COVID-19 time-series data from the state of Arkansas. The green curves are the centroids Ccj, and the red curves are the best-fit regression functions PFM_j for each cluster C_j

For the given conditions, the polynomial regression clustering method tends to produce 5-9 clusters on each instance of the training data. An example of polynomial clustering produced from the Arkansas data is shown in Fig. 4, and Fig. 3 displays entire time-series. The green curves plot the centroid functions Ccj, and the red curves plot the best-fit regression functions PFM_j for each cluster C_j. Upon visual inspection, the centroids give us a general idea of the different kinds of features exhibited by small slices in the time-series data used in this study.

Though samples in some clusters appear extremely similar to each other (row 3, column 1), other clusters, such as the one depicted in row 1, column 2, display concerning behavior. While this cluster's centroid is decreasing, many of its samples are increasing with time. This is due to a limitation of the distance function used to cluster these data. The DyCPR algorithm operates by minimizing the cosine distance between centroids and samples, but it does not take into account the behavior of polynomials as time increases without bound. It is possible that this behavior decreases prediction accuracy with this model.

One modification that could improve the quality of

clusters would involve evaluating derivatives of samples and clusters. Before sample PF_i is grouped with its most similar cluster C_j, evaluate the derivatives of sample PF_i and centroid Cc_j at time n. If $PF_i'(n) * Cc_j'(n) < 0$, then the sample and centroid have differing limits as time increases without bound, so we create a new cluster from PF_i. This small modification to DyCPR would take into account a polynomial curve feature that is easily identified on visual inspection, but impossible to evaluate using only the cosine distance function. The resulting clusters would then have greater similarity among their samples, and this in turn could increase the accuracy of prediction with this model.

V. Results

After training, the model is tested using the test set of time-series slices of length $n+1$. Given the first n data points from a slice spanning time 1 to $n+1$, the model uses polynomial regression to find the most similar cluster C_j, then uses PFM_j corrected with y_n to predict the value y_{n+1} (Fig. 5) (see Step 3 in Section III for detailed description).

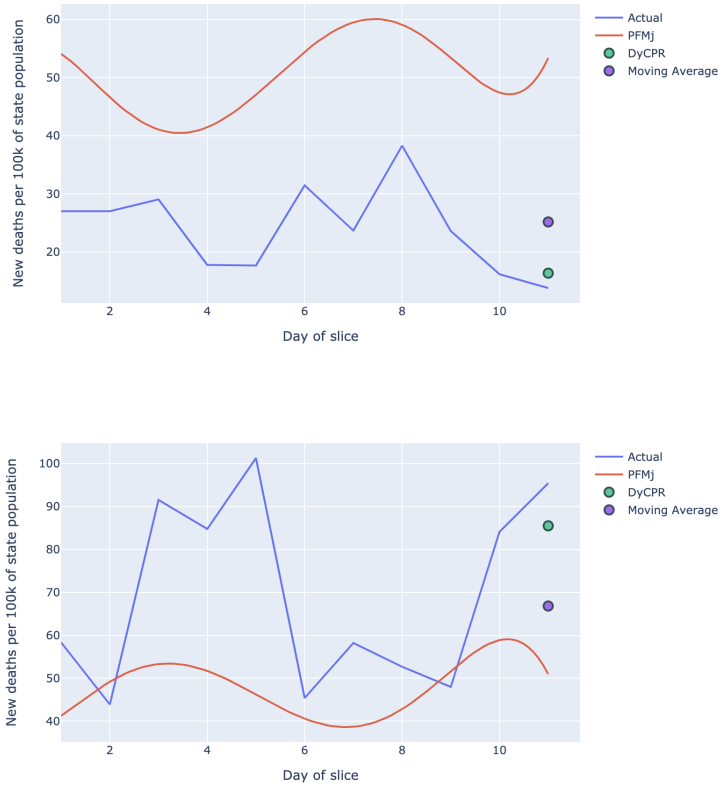

Fig. 5. Two predictions made by DyCPR on the test data compared to moving average, for the Arkansas daily COVID-19 cases time-series data.

In addition to the DyCPR predictions, an n-day moving average algorithm was used to predict values on the same test samples. The root mean squared error (RMSE) was used to evaluate the accuracy of both the DyCPR and moving average predictions, and the average RMSE over 10 trials for each state's data were tabulated (Fig. 6). The results of the study show that for 18 US states, DyCPR was more accurate than moving average over 10 trials in predicting the number of new COVID-19 cases on day 11, given the counts for the previous 10 days.

VI. Conclusion

In this study, DyPCR made more accurate predictions than moving average on COVID-19 time-series data for 18 out of 51 states. I expected DyCPR to out-perform moving average on prediction for states with more chaotic data, however, it seems that DyCPR was actually more accurate for making predictions on stabler time-series data. For example, DyCPR was more accurate in predicting values for New York and California, states which have exhibited relatively stable growth in daily increases of reported COVID-19 cases. I did not explore how the amount of chaos in the data affects prediction accuracy, but future works could expand upon this. Using Lyapunov exponents could give a measure of chaos to a given set of time-series data. Then, different prediction models could be assessed to determine a measurable rule for choosing the right model based on the amount of chaos in the data.

As the COVID-19 pandemic continues to claim more lives every day, government officials and healthcare providers could use prediction models like DyCPR for decision-making. For example, future studies could be performed on a much smaller scale–on the city or county level–to determine the staff schedule or medical supplies needed in a hospital on a given day.

References

[1] *Population Estimates by Age (18+): July 1, 2019*, United States Census Bureau, 2019. (dataset). https://www.census.gov/data/tables/time-series/demo/popest/2010s-national-detail.html. (accessed Dec. 9, 2020).

[2] *State-by-state Data - A list of COVID data by state*, The Atlantic Monthly Group, 2020. (dataset). https://covidtracking.com/data. (accessed Dec. 11, 2020).

[3] "WHO Coronavirus Disease (COVID-19) Dashboard.", World Health Organization. https://covid19.who.int (accessed Dec. 14, 2020)

[4] Widiputra H., Kho H., Lukas, Pears R., Kasabov N. (2009) A Novel Evolving Clustering Algorithm with Polynomial Regression for Chaotic Time-Series Prediction. In: Leung C.S., Lee M., Chan J.H. (eds) Neural Information Processing. ICONIP 2009. Lecture Notes in Computer Science, vol 5864. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10684-2_13

State	Data Range	DyCPR RMSE (10 trial avg)	Moving Avg RMSE (10 trial avg)	Higher Accuracy
Alaska	3-168	10.72	12.49	DyCPR
Alabama	2-125	18.86	15.43	Moving Average
Arkansas	11-122	16.6	16.23	Moving Average
Arizona	1-219	25.43	18.45	Moving Average
California	5-116	6.76	8.78	DyCPR
Colorado	4-144	10.3	11.96	DyCPR
Connecticut	0-287	33.83	23.23	Moving Average
Delaware	1-51	5.3	5.01	Moving Average
District of Columbia	2-164	16.71	17.0	DyCPR
Florida	4-98	11.27	9.32	Moving Average
Georgia	7-75	8.85	8.96	DyCPR
Hawaii	0-36	4.59	4.58	Moving Average
Iowa	7-181	20.32	21.33	DyCPR
Idaho	5-172	18.23	17.23	Moving Average
Illinois	5-157	13.58	13.19	Moving Average
Indiana	5-164	10.68	13.04	DyCPR
Kansas	22-341	21.02	28.04	DyCPR
Kentucky	6-163	17.0	14.22	Moving Average
Louisiana	1-150	30.26	23.16	Moving Average
Massachusetts	2-121	9.02	9.43	DyCPR
Maryland	5-81	5.95	6.25	DyCPR
Maine	0-39	3.73	3.54	Moving Average
Michigan	4-222	21.45	17.84	Moving Average
Minnesota	5-209	18.07	17.92	Moving Average
Missouri	1-134	15.18	18.08	DyCPR
Mississippi	6-121	17.59	16.75	Moving Average
Montana	4-196	25.14	22.77	Moving Average
North Carolina	6-98	11.83	10.35	Moving Average
North Dakota	5-401	32.55	33.53	DyCPR
Nebraska	5-236	22.35	21.99	Moving Average
New Hampshire	0-108	8.6	8.37	Moving Average
New Jersey	0-87	7.16	6.83	Moving Average
New Mexico	2-227	13.37	18.27	DyCPR
Nevada	5-134	14.31	13.23	Moving Average
New York	2-74	3.36	4.57	DyCPR
Ohio	6-283	18.74	18.96	DyCPR
Oklahoma	5-209	25.76	19.62	Moving Average
Oregon	3-65	4.99	4.87	Moving Average
Pennsylvania	0-127	7.57	9.93	DyCPR
Rhode Island	2-212	22.97	20.48	Moving Average
South Carolina	7-88	11.63	9.56	Moving Average
South Dakota	3-321	37.47	34.34	Moving Average
Tennessee	12-155	28.26	20.37	Moving Average
Texas	6-83	11.64	12.62	DyCPR
Utah	10-271	30.55	21.44	Moving Average
Virginia	5-66	6.05	6.15	DyCPR
Vermont	0-44	4.1	3.34	Moving Average
Washington	0-117	16.43	13.15	Moving Average
Wisconsin	5-187	20.96	20.33	Moving Average
West Virginia	0-100	9.46	8.17	Moving Average
Wyoming	2-284	53.64	32.92	Moving Average

Fig. 6. Results from DyCPR and moving average predictions, over 10 trials for each US state. For 18 states, DyCPR was more accurate than moving average in predicting the value on day 11, given the previous 10 days.