DOI: 10.1002/pra2.355

POSTERS

WILEY

Through the magnifying glass: Exploring aggregations of COVID-19 datasets by county, state, and taxonomies of U.S. regions

Yi-Yun Cheng | **Bertram Ludäscher**

School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, Illinois

Correspondence
Yi-Yun Cheng, School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL.
Email: yiyunyc2@illinois.edu

Abstract

In this preliminary study, we investigate the case of COVID-19 United States confirmed cases datasets, and perform experiments with aggregations of data by county, state, and different taxonomies for U.S. regions. The overarching goals of this study is to uncover potential data quality issues due to different levels of geospatial aggregation of data.

KEYWORDS

COVID-19, data aggregations, data quality, taxonomies

1 | INTRODUCTION

As the COVID-19 pandemic progresses, discussions about data quality issues of COVID-19 datasets abound. For example, counts of infected, tested, and recovered persons are susceptible to misrepresentation due to the choices of empirical case counting, and variables unaccounted for (Maier & Brockmann, 2020). While data quality issues continue to be of great importance and have led to controversies (Ioannidis, 2020), research about aggregation of data based on different taxonomies in the context of COVID-19 can be further examined. In the context of this preliminary study, data quality issues refer to the presence of overlapping, contradicting, and inconsistent data at the instance-level (Rahm & Do, 2000). We explore the different aggregation units by county, state, and region of the COVID-19 datasets. Regional aggregation may be further complicated by using alternative regional groupings, based on different taxonomies. We hope to initiate conversations on possible new data quality issues brought forth by the geographic granularity and taxonomy of datasets.

2 | METHOD

2.1 | Data collection

We obtained the COVID-19 United States confirmed cases datasets from the Johns Hopkins University (JHU) repository.[1]We collected the datasets until May 30, 2020. Two datasets are used in this study:

1. **Time_series_COVID-19_confirmed_US** (*time series dataset*): The time series dataset documents the number of confirmed cases by *county-level* in the United States from Day 1 (January 22, 2020) when the first case of COVID-19 was confirmed to present day. In this study, we focus only on the contiguous U.S. (Lower 48 states and the District of Columbia).
2. **COVID-19_daily_reports_US** (*overview dataset*): The overview dataset contains information by *state-level* of the total number of confirmed, deaths, recovered, etc., as of a

83rd Annual Meeting of the Association for Information Science & Technology October 25-29, 2020.

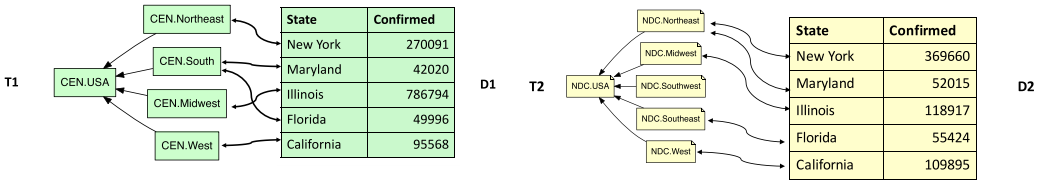

FIGURE 1 Example of transforming a small set of the overview dataset into regional-level data

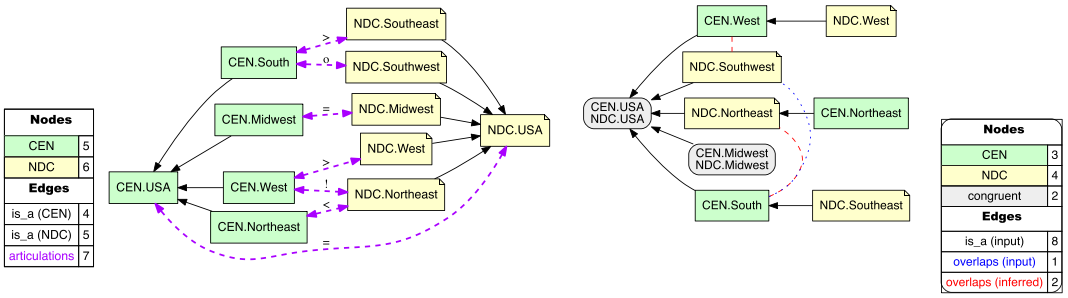

FIGURE 2 The input alignments of T_{CEN}, T_{NDC}, and the relations (left); the output merged solution (right)

particular day. For this study, we use the overview dataset of May 30, 2020.

2.2 | Taxonomies

Two taxonomies are used in this study to create a new, experimental use case, where each taxonomy represents a different grouping of US states into regions.

1. **Census Bureau (T_{CEN}):** the Census Bureau divides the contiguous U.S. into four regions, namely Midwest, Northeast, South, and West.
2. **National Diversity Council (T_{NDC}):** the national diversity council divides the contiguous U.S. into five regions—Midwest, Northeast, Southeast, Southwest, West.

2.3 | Constructing the experimental datasets

We transform the time series dataset and the overview dataset into regional-level datasets by linking the entities in each dataset with the two geographic taxonomies. Figure 1 shows how the overview dataset is converted into two dataset D_1 and D_2: D_1 uses T_{CEN}, while D_2 uses T_{NDC}.

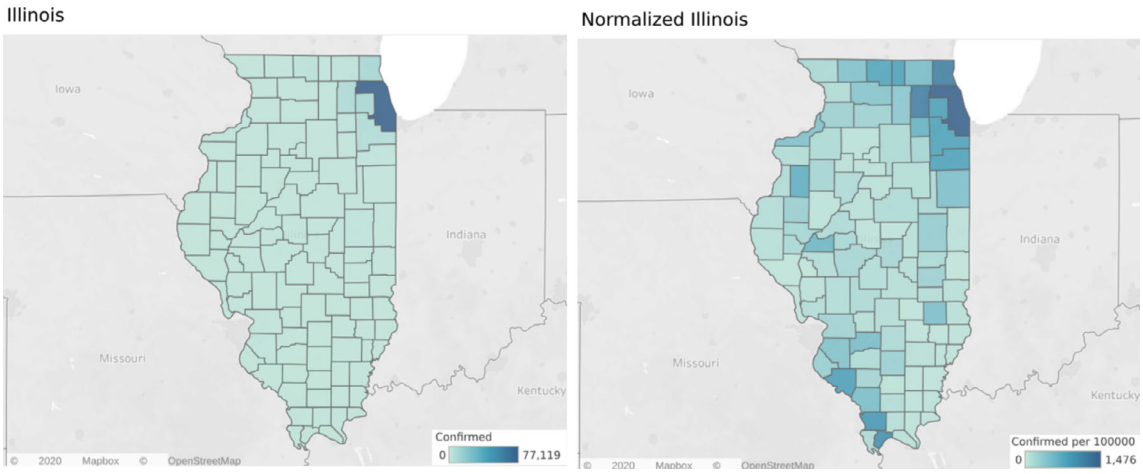

FIGURE 3 Illinois confirmed cases by county. Absolute (left); normalized (right)

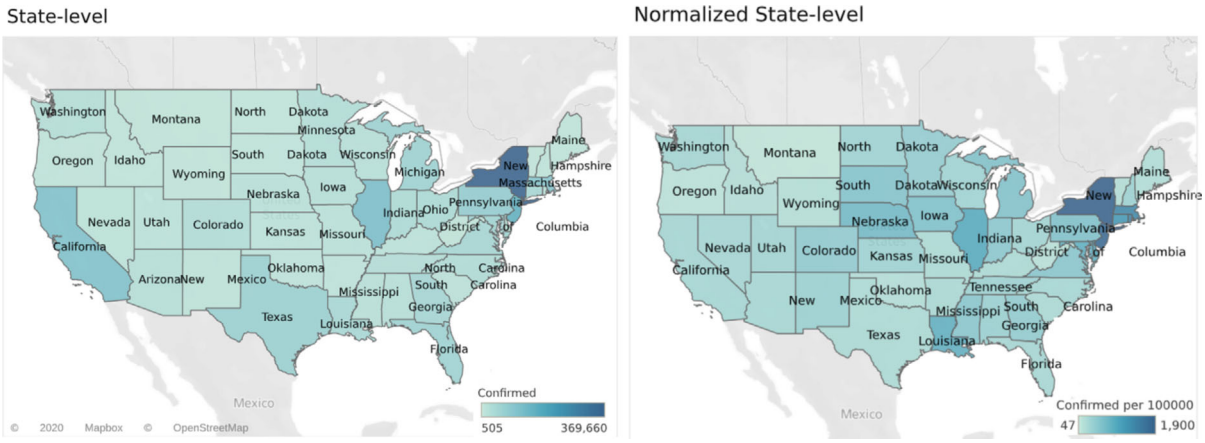

FIGURE 4 Confirmed cases by state. Absolute (left); normalized (right)

2.4 | Reconciliation of taxonomies

T_{CEN} and T_{NDC} are aligned and reconciled into a combined or "merged" taxonomy via a logic-based taxonomy alignment approach (Cheng et al., 2017). The method uses a qualitative reasoning approach (RCC-5), in which concepts in T_{CEN} are mapped to T_{NDC} using one of five base relations: equivalence, overlap, disjointness, inclusion, and inverse inclusion. The two taxonomies, when aligned via the RCC-5 relations, then form one or more merged solutions. In this study, the two taxonomies yield a single, unique solution (Figure 2).

3 | RESULTS AND DISCUSSION

We demonstrate the differences between granularities on geographic regions of the U.S., starting from the finer-grained county-level analysis, state-level analysis, to the most coarse-grained regional-level analysis. We also explore the differences in using the absolute counts of the confirmed COVID-19 cases (*absolute*) and the normalized counts (i.e., per 100,000 people) in a particular area (*normalized*).

3.1 | County-level

Zooming into Illinois counties, there is a notable difference in the absolute and normalized total cases. While Cook county remains top ranked in both the absolute ($n = 77{,}119$) and the normalized ($n = 1{,}476$), the ranking shifts for the remainder counties. We see drastic changes between the two visualizations in Figure 3: in the absolute counts only Cook County is particularly hard hit (dark blue), while the normalized shows additional "hot spots" for example, in southern counties and counties neighboring other states.

3.2 | State-level

Figure 4 shows the confirmed cases by state. Looking at absolute numbers, one might be misled to think that apart from New York ($n = 369{,}660$) and New Jersey ($n = 159{,}608$), things are mostly under control. But the normalized view ranks states in a different order: New York ($n = 1{,}900$), New Jersey ($n = 1{,}796$), Rhode Island ($n = 1{,}399$), Massachusetts ($n = 1{,}397$), and District of Columbia ($n = 1{,}235$) are all hit heavily as of May 30, 2020.

3.3 | US-region level

Figure 5 depicts cases by U.S. region: in T_{CEN}, the region with most cases is the Northeast ($n = 765{,}858$), followed by South ($n = 437{,}238$), Midwest ($n = 351{,}201$), and West ($n = 210{,}651$). Northeast ($n = 1{,}368$) is still the most severe in the normalized view, but the ranking shifts for Midwest ($n = 514$), and South ($n = 348$). West is still the least severe ($n = 276$).

T_{NDC} also shows that Northeast is still the most severe ($n = 836{,}012$), followed by Midwest ($n = 351{,}201$), Southeast ($n = 297{,}991$), West ($n = 183{,}769$), and Southwest ($n = 95{,}975$). The normalized shows the same ranking: Northeast ($n = 1{,}312$), Midwest ($n = 513$), Southeast ($n = 350$), West ($n = 275$), Southwest ($n = 226$).

Comparing across normalized T_{CEN} and T_{NDC}, it appears that the Southwest is less impacted, since the Southeast is its own region. Not surprisingly, when comparing data across levels (county, state, region), differences

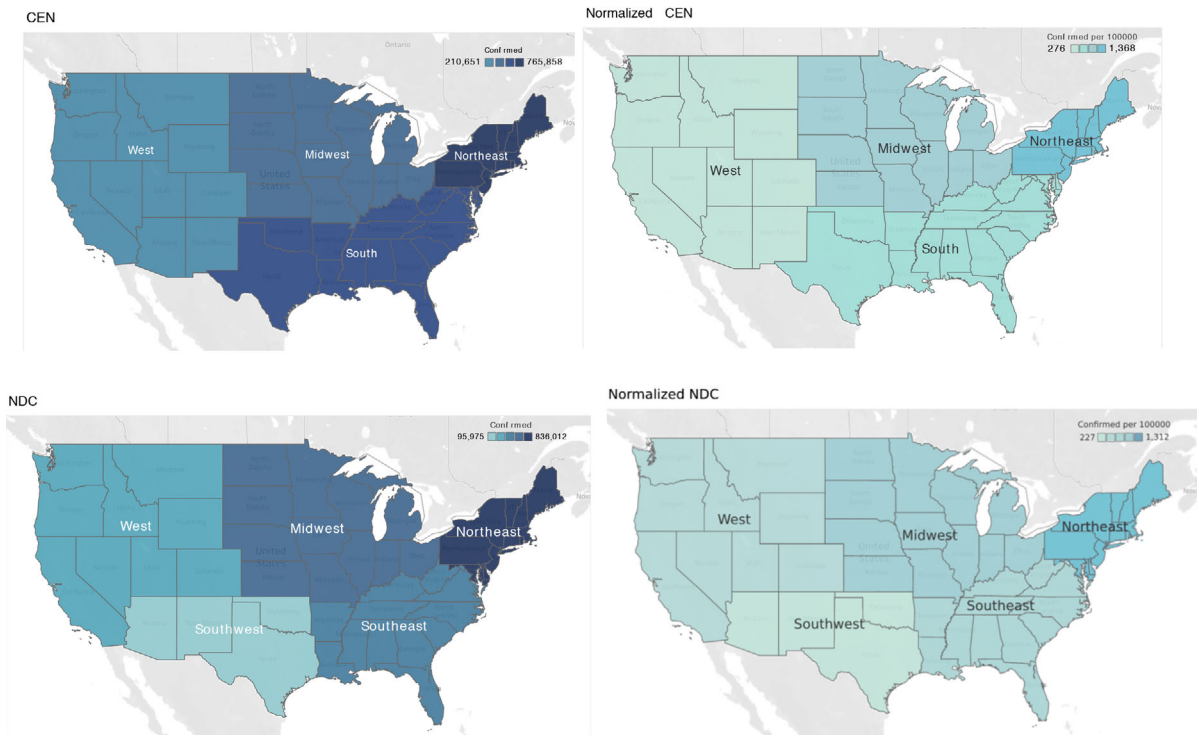

FIGURE 5 Confirmed cases by region and taxonomy. CEN regions: West, Midwest, South, Northeast; NDC regions: West, Midwest, Southwest, Southeast, Northeast. Absolute (left); normalized (right). T_{CEN} (top); T_{NDC} (down)

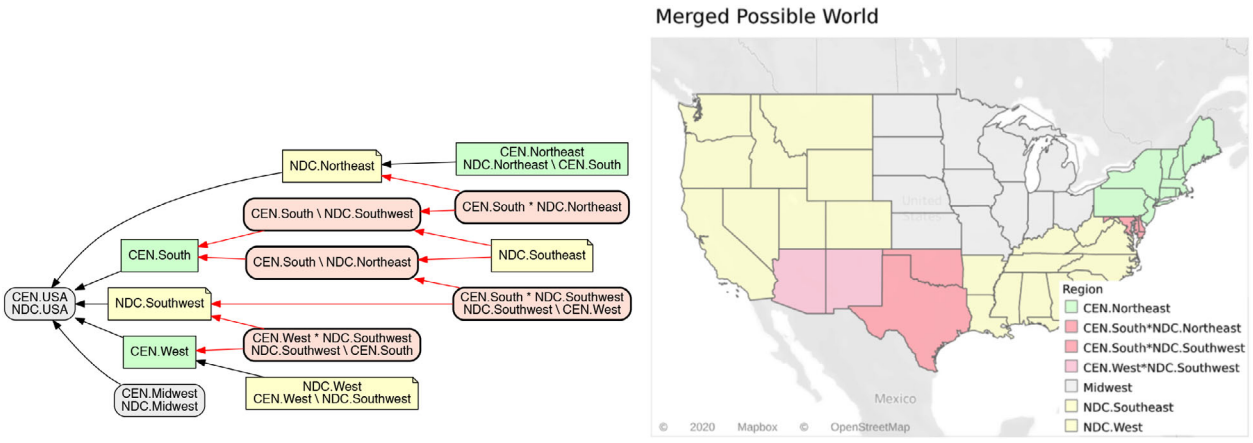

FIGURE 6 Merged taxonomy for T_{CEN} and T_{NDC} and its corresponding map view

tend to appear more "washed out" at the coarser levels of aggregation.

3.4 | Taxonomic views

Reconciling the two taxonomies T_{CEN} and T_{NDC} returns the merged view shown in Figure 6, where concepts from T_{CEN} and T_{NDC} are preserved, new regions (in pink) emerge to show where the two taxonomies differ. At the leaf-level, there are seven nodes in total, each corresponds to a region in the map view of the merged taxonomy.

Figure 7 shows how the merged taxonomies can be used in datasets to show different representations from T_{CEN} or T_{NDC}. The absolute numbers show CEN.Northeast, Midwest, and NDC.Southeast as top three, but most of the regions are also severely impacted. However,

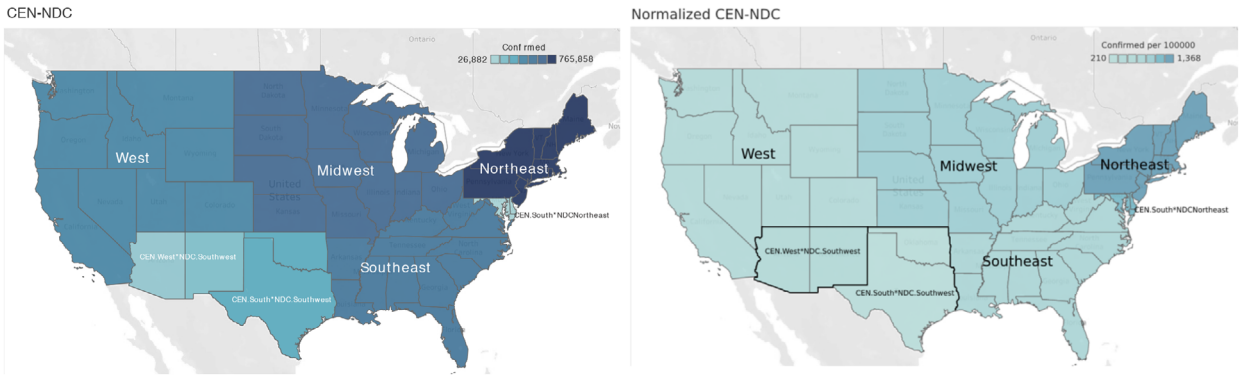

FIGURE 7 Confirmed cases by merged taxonomy. Absolute (left); normalized (right)

the normalized view suggests CEN.Northeast (n = 1,368), CEN.South*NDC.Northeast (n = 908), and Midwest (n = 514) as the top three regions, with the other four regions having less than 350 cases per 100,000 people.

4 | CONCLUSIONS

In this study, we have examined COVID-19 datasets (confirmed cases) at different geographic resolutions. While many data quality issues are already known, *additional* problems may arise when employing different geo-taxonomies for coarse-grained regions. Analysis of geospatial data should usually be done at the finest (e.g., county-level) resolution available. However, coarser-grained aggregations are frequently used by the media to report events and the "big picture" (e.g. "Midwest is the new epicenter of COVID-19", "South is opening up soon."). The results of this paper suggest that aggregation at coarse-grained levels have to be treated with great caution: (a) aggregation loses important detail and may underestimate and/or overestimate the severity of the virus spread; (b) different aggregations due to alternative taxonomies (regional groupings) may create additional confusion; and (c) reconciliation of taxonomies may be useful prior to merging datasets.

ENDNOTE

[1] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

REFERENCE

Cheng, Y. Y., Franz, N., Schneider, J., Yu, S., Rodenhausen, T., & Ludäscher, B. (2017). Agreeing to disagree: Reconciling conflicting taxonomic views using a logic-based approach. *Proceedings of the Association for Information Science and Technology*, *54*(1), 46–56.

Ioannidis, J. P. (2020). A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data. *STAT*, *17*. Retrieved from: https://www.statnews.com/2020/03/17/a-fiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-without-reliable-data/.

Maier, B. F., & Brockmann, D. (2020). Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, *368*(6492), 742–746.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, *23*(4), 3–13.

How to cite this article: Cheng Y-Y, Ludäscher B. Through the magnifying glass: Exploring aggregations of COVID-19 datasets by county, state, and taxonomies of U.S. regions. *Proc Assoc Inf Sci Technol*. 2020;57:e355. https://doi.org/10.1002/pra2.355