

How to provide light to COVID data by means of FCA^{*}

Domingo López-Rodríguez^[0000-0002-0172-1585], Pablo Cordero^[0000-0002-5506-6467], Manuel Enciso^[0000-0002-0531-4055], and Ángel Mora^[0000-0003-4548-8030]

Universidad de Málaga, Málaga, Spain
{dominlopez,pcordero,enciso,amora}@uma.es

Abstract. COVID data are usually presented in a non-structured format and mainly focused on healthy issues (incidence, mortality, etc). At the same time, Governments have designed a set of measures to deal with the Pandemic. In addition, several institutions have studied the economical effects of the situation in each country. In this work, we combine these three data sources and illustrate how Formal Concept Analysis can become a useful tool to discover relationships among these three views of the situation: health, politics and economy. Our aim is to provide an implication-driven approach to discover knowledge behind the data.

Keywords: Formal Concept Analysis · Implications · Covid.

1 Introduction

During 2020 and so far in 2021, the world has been paralysed by the pandemic. The devastating effects of COVID-19 in terms of loss of human life, the limitations in every country worldwide and the economic downturn constitute a dramatic landmark in the history of humankind.

From the very beginning, individual researchers, universities and institutions began to accumulate data on the virus incidence oriented to measure the stress of the health infrastructures and the impact of the Pandemic on mortality. Simultaneously, governments around the world have implemented several restrictions to minimise the incidence of the virus. Each country has taken different measures depending on their situation and, as usual, taking into account the political effects of these measures on the citizens [15].

Data scientists have conducted extensive, fast and effective work, showing the power of these methods and tools and providing some valuable guidelines to support the fight against the virus and check its evolution. Most of these

* Partially supported by the projects TIN2017-89023-P (Spanish Ministry of Economy and Competitiveness, including FEDER funds), PGC2018-095869-B-I00 (Spanish Ministry of Science and Innovation), and the Junta de Andalucía projects UMA2018-FEDERJA-001, UMA18-FEDERJA-158 and UMA-CEIATECH-24 co-funded by the European Regional Development Fund.

works have been oriented for diagnoses (using image processing or tracking techniques) or prediction and forecast (time series analysis or machine learning methods) [8]. Other authors have developed complex mathematical models to explain the virus behaviour (nonlinear fractional-order differential equations [1], Chaos theory [14], etc.).

To the best of our knowledge, easy and direct relationships looking at the evolution curve of the number of infected persons and deaths can be extracted if we insert in the evolution curve the major measures taken by governments such as the total closure of companies and the confinement of people to their homes. These classical and quantitative relationships can be considered a good but basic approach. We are not aware of studies that allow studying all the elements involved in the three-dimensional data (health, politics and economy). We need to address the following problem: how to relate all the measures taken to a greater or lesser extent with the changes in the evolution of the number of infected persons, deaths, admissions to intensive care units and the interaction of all this with the economic downturn.

The first step is to build a comprehensive dataset from reliable sources.

Regarding the disease dimension and its impact, there exist many sources collecting up-to-date data. Mainly, we have used the data from the European Centre for Disease Prevention and Control (ECDC), an agency of the European Union¹ because of its trustworthiness. To complete our study, we have integrated economy and politics measures data. On the one hand, we have collected Eurostat data about the Economy. We have used the repository that this institution has mainly created to measure the impact of the crisis in Europe². On the other hand, we have collected the measures designed for all the EU countries in the different waves of the pandemic as well as the number of cases and deceases, hospital occupancy and other related variables³.

Formal concept analysis [4] (FCA) has shown to be a strong and solid area to deal with the complete data life-cycle, from the data extraction, knowledge representation and management. However, this solid development has still to be fruitful in real applications, as others areas are. FCA has also been used to address COVID Pandemic. In [6], the authors built a concept lattice to navigate, providing successful search in COVID resource platforms. In [2] the author characterised a set of attributes of the vaccines and created a concept lattice to provide a hierarchical landscape of the current status of phase 3 vaccines. However, few works show FCA's power to design a data-driven approach for this issue. In this work, we propose to extract valuable knowledge from public data using the so-called implications (if-then rules). These rules can be used to discover the interesting relationship among the three dimensions involved in the Pandemic. Such relationships are not merely cause-effect rules, and they provide a complete portrait of the semantic behind the data.

¹ <https://www.ecdc.europa.eu/en>

² <https://ec.europa.eu/eurostat/web/COVID-19/data>

³ <https://www.ecdc.europa.eu/en/COVID-19/data>

In this work, we are using `fcaR` software developed in our research group, *Malaga-FCA-group*, using R language. We propose this tool as a useful tool to approach real problems in FCA and illustrate the benefits of this solid framework. See <https://github.com/Malaga-FCA-group/fcaR> for more information.

The paper is organised as follows. In section 2 we describe how to address the import, processing and data curation in this real problem and how to integrate such data to be further used with FCA methods. In Section 3 we explore the discovery of knowledge extracted from the data (using implications) of COVID to establish how to act better in the future to reduce the incidence of the virus and minimise economic losses. We aim to illustrate how implications can be considered a suitable tool to help in decision-making tasks. Section 4 proposes some conclusions and future works.

2 Combination of open data into one integrated dataset

As we have explained in the introduction, we have downloaded the open data regarding health and measures from UE institutions availables in the aforementioned sources. ECDC publishes weekly updates. Briefly, we have collected data about the following information: data on 14-day notification rate of new COVID-19 cases and deaths, data on hospital and ICU admission rates and occupancy for COVID-19, data on country response measures to COVID-19, data on Harmonised unemployment rates and data on Gross Domestic Product (GDP)

In this first study, we also have introduced the time dimension to infer conclusions about the Pandemic evolution. Thus, we take the second and third quarters (Q2, Q3) of 2020 as the study periods ⁴, because of the harshness of the incidence of the virus during these months and because all the EU countries in these two periods have presented significant differences in their situation.

On the other hand, these variables are both numerical and dates, then we have opted to perform conceptual scaling after normalizing the attributes. We summarize the data pre-processing we have done to achieve the formal context being the target of the application of the FCA methods:

- Each measure adopted in the EU countries in one specific quarter is stored in one column. For instance, `MEAS_Q2_MasksMandatoryClosedSpaces` means that in the second quarter (Q2), masks were mandatory in all closed spaces.
- We use the NOT token to show the absence of a given measure⁵. For instance, `MEAS_NOT_Q2_AdaptationOfWorkplace` means that in the second quarter, the country did not take any action on the adaptation of workplaces.
- For each quarter (Q2 and Q3), we have discretized the COVID data (deaths and cumulative cases in percent wrt population), ICU and hospitalization data, and excess mortality data. For instance, `COV_Q2_cases_low` means that

⁴ Remark that we use somehow a temporal information, but we do not make use of temporal FCA [16] to ease a simpler management in the future.

⁵ Note that in this paper we do not introduce the generalization of FCA to consider negative attributes [12,10]. We will address this issue in future works.

in quarter two the country had a low level of COVID confirmed cases, `COV_Q3_ExcessMortalityChange_decreased` means that in Q3 was detected a decrease in the excess of mortality in the country.

- For economical data, we have collected information about **changes** in *Unemployment rates* and *GDP*.

3 Knowledge representation and its interpretation

After the pre-processing task, we got an adequate binary formal context to be explored with FCA. This formal context has 175 attributes and 25 objects. The dataset has been imported in R using the `fcaR` package. From this point, we performed the following steps:

- First, we remove *constant* columns, corresponding to attributes (measures or health or economic parameters) that are present either in none or in all the objects (the 25 countries in the EU).
- Then, we perform clarification on the resulting formal context to identify equivalent attributes, that is, attributes a, b such that $a \Rightarrow b$ and $b \Rightarrow a$.
- We compute the Duquenne-Guigues basis of implications and inspect them to get a complete knowledge representation of the data allowing further symbolic manipulation of the coronavirus situation in the EU countries.
- In the last step, we apply the Simplification Logic for implications [13] to remove *attribute* redundancies in the above basis, keeping an equivalent set of implications with lower size (number of attributes in each implication).

The Duquenne-Guigues basis of this formal context has 4086 implications. The average size of its LHS and RHS is 11.297 and 1.594 attributes, respectively. After using the simplification logic to remove redundancies, the resulting implication set has 4.472 and 1.188 attributes in the average size in LHS and RHS. The decrease of the implication size eases its reading and interpretation.

In the next subsections, we present the main findings obtained from the simplified basis of implications. Our goal is to illustrate how these implications can be used to get a useful interpretation of the model.

3.1 Common strategies and parameters

Some measures weren't taken by any country in the management of the pandemic. Their interpretation depends on the measure or effect and provides useful information even though its uniform absence seems to induce no information.

For instance, in the 3rd quarter, no country imposed a ban on all events or restrictions on private gatherings or closed restaurants, hotels and entertainment venues. Concerning economic parameters, the unemployment rate wasn't stable in both quarters. It remarks that the strong lockdown measures were not needed in the Q3 period.

It can be observed that the GDP in the second quarter decreased in all countries whereas it increased in the third quarter (summer). And, since there is

a common value for all the objects (countries), this progress occurs irrespective of the restriction measures taken.

3.2 Equivalent attributes

We named equivalent attributes those ones that play the same role, and they can be characterized by using attribute reduction and clarification methods [7].

The attributes `NOT_Q2_EntertainmentVenues` and `Q3_StayHomeRiskG` are found to be equivalent. This means that every country that did not close entertainment venues in the 2nd quarter (Q2) of 2020, in the 3rd quarter (Q3), they had to issue a recommendation to stay home for risk groups (older adults, disabled people or people with other pathologies).

Also, having a high number of deaths in Q2 was equivalent to a high number of deaths in Q3. This is interesting since no matter the measures adopted if the pandemic has led to a high number of deaths during the first stage, in the summer, the number of deaths remains at a high level, and it was also greater than Europe's average.

A high hospitalization occupancy ratio is shown to be equivalent to the ICU occupation ratio during Q2. This data illustrates that the evolution of patients admitted in the hospitals is not very promising. This is well-known for the health experts that insistently alert us about the risks of this disease, and the FCA analysis also confirms it.

3.3 Implications provide useful interpretation

To illustrate how to get useful knowledge from the implications, we have selected some implications whose interpretation may provide a deeper insight on the pandemic management strategies and the consequences of taking restriction measures. Since the implication set is large, we only show some easily interpretable and with a low number of attributes, and with positive support. In the following, we will simplify attributes' names to allow for a better readability:

- Some implications on the influence of Q2 measures in Q2 COVID cases or deaths (short-term effects in health):

<code>NOT_Q2_MassGather50</code>	\Rightarrow	<code>Q2_cases</code> are high
<code>Q2_AdaptationOfWorkplace</code>	\Rightarrow	<code>Q2_cases</code> are medium
<code>Q2_RestaurantsCafes, Q2_MasksMandatoryAllSpaces</code>	\Rightarrow	<code>Q2_deaths</code> are medium

These implications can be interpreted as follows: the countries that didn't constraint mass gatherings of more than 50 people in the Q2 ended this period with a number of COVID cases greater than Europe's average. By contrast, the countries that promoted adaptation of workplaces ended with a number of cases equivalent to the average. However, such promising impact of the virus can also be provided by other measures, and the same situation appears in the countries that did close restaurants and where masks were mandatory in all spaces.

- Influence of Q2 measures in Q2 economy (short-term effects in economy):

NOT_Q2_WorkplaceClosures \Rightarrow Q2_UnemploymentChange is increased
 NOT_Q2_StayHomeOrder \Rightarrow Q2_UnemploymentChange is increased

These implications have a potentially *counterintuitive* interpretation. Countries that did not close workplaces or didn't order a strict lockdown had their unemployment rate increased; that is, not taking those measures didn't prevent a rise in the unemployment rates. It is important to highlight this relationship that has been discovered with the help of FCA tools.

- Influence of Q2 measures in Q3 COVID (mid-term effects in health):

NOT_Q2_StayHomeRiskG, NOT_Q3_MasksVoluntaryAllSpaces
 \Rightarrow Q3_ChangeInHospital is increased

Those countries that didn't recommend lockdown for risk groups in Q2 and didn't impose masks in all spaces in Q3, had their number of hospitalizations rise from Q2 to Q3.

- Influence of Q2 measures in Q3 economy (mid-term effects in economy):

NOT_Q2_BanOnAllEvents \Rightarrow Q3_UnemploymentChange is decreased
 NOT_Q2_StayHomeOrder \Rightarrow Q3_UnemploymentChange is decreased
 Q2_PrivateGatheringRestrictions \Rightarrow Q3_UnemploymentChange is decreased

The unemployment decreased in those countries that, in Q2, did not ban all events, nor ordered a strict lockdown nor had restrictions on private gatherings.

- Influence of Q3 measures in Q3 COVID:

NOT_Q2_StayHomeRiskG, NOT_Q3_MasksVoluntaryAllSpaces
 \Rightarrow Q3_ChangeInHospital is increased
 NOT_Q2_StayHomeRiskG, NOT_Q3_MasksMandatoryAllSpaces
 \Rightarrow Q3_ChangeInHospital is increased
 Q3_StayHomeRiskG
 \Rightarrow Q3_inICU is medium, Q3_ChangeInHospital is decreased,
 Q3_ExcessMortality is equivalent

The first two implications tell us that not recommending risk groups to stay home and not suggesting the use of masks in all spaces led to an increase in hospitalizations during summer. The countries that recommended risks groups to stay home ended with an average ICU occupancy rate, a decrease in hospitalizations and a normalization of the mortality excess (that is, the number of deaths in the country is equivalent to the previous years).

3.4 Closure of attributes to find common properties

To complete the information, we have studied the closure of some interesting attributes. In particular, we want to answer the question ‘*What had in common the countries with a low and a high hospital occupancy rate in the second quarter?*’

To answer this question, we will compute the closures of the corresponding attributes (`Q2_inHospital` is low and `Q2_inHospital` is high).

- The closure of the first attribute is the following `Q2_ClosHigh`, `Q2_ClosPrim`, `Q2_MassGather50`, `Q2_MassGatherAll`, `Q2_OutdoorOver1000`, `Q2_EntertainmentVenues`, `NOT_Q2_ClosureOfPublicTransport`, `NOT_Q3_MasksMandatoryAllSpaces`, `NOT_Q3_IndoorOver50`, `NOT_Q3_OutdoorOver50`, `NOT_Q3_Teleworking`.

This means that countries that ended Q2 quarter with low occupancy had closed high and primary schools and imposed restrictions on mass gatherings and outdoor. However, in the third quarter, they could relax some restrictions on gatherings because of the low occupancy.

- The high occupancy closure provides the following set of attributes: `Q2_ClosSec`, `NOT_Q2_HotelsOtherAccommodation`, `NOT_Q2_ClosureOfPublicTransport`, `NOT_Q3_Teleworking`, `Q2_ExcessMortality is more`, `Q3_ExcessMortalityChange is decreased`.

This result means that they didn’t have a common strategy and the only interesting consequence is that in Q3, the mortality excess was decreased.

4 Conclusions

In this work, we have used FCA to address the problem to illustrate the relationships among the government measures, the economic impact and the health situation in the Pandemic. We have combined real public data from all countries in the EU stored in different data sets, establishing a data pre-processing task. We have made a knowledge representation through the implications using the `fcaR` package. Reducing the original Duquenne-Guigues basis using the Simplification logic eases a further examination of the set of implications, allowing to deduce valuable insights about the COVID disease, its impact and the measures that can be used to deal with the virus.

As future works, we plan to collect more data. We intend to consider all the countries worldwide and collect more attributes for the evolution of the data in 2021. A more comprehensive conceptual scaling should be performed to manage such large dataset. In addition, it is of interest the use of generalized attributes [11] to handle the increasing number of attributes.

To provide more information about the new data, we will also build the concept lattice that provides other helpful information and a graphical representation of the information, essential for further navigation among the closure set of attributes. Since the number of concepts increases when more data were considered, we will need a filtering method to select those which are most likely to provide promising information. Stable concepts may provide more insights in

this delicate issue [9]. Moreover, we have to study the methods and techniques to develop some kind of conceptual exploration [5].

Finally, we will also address the negation issue (the relation between contrary attributes) by means of mixed (negative and positive) attributes [3].

References

1. Idris Ahmed, Isa Abdullahi Baba, Abdullahi Yusuf, Poom Kumam, and Wiyada Kumam. Analysis of caputo fractional-order model for covid-19 with lockdown. *Advances in Difference Equations*, 2020(1):1–14, 2020.
2. Javier Dario Burgos-Salcedo. A comparative analysis of clinical stage 3 covid-19 vaccines using knowledge representation. *medRxiv*, 2021.
3. P. Cordero, M. Enciso, A. Mora, and J. M. Rodríguez-Jiménez. Inference of mixed information in formal concept analysis. *Studies in Computational Intelligence*, 796:81–87, 2019.
4. B. Ganter and R. Wille. *Formal Concept Analysis. Mathematical Foundations*. Springer, Berlin, 1996.
5. Bernhard Ganter and Sergei A. Obiedkov. *Conceptual Exploration*. Springer, 2016.
6. Fei Hao and Doo-Soon Park. Conavigator: A framework of fca-based novel coronavirus covid-19 domain knowledge navigation. *Human-centric Computing and Information Sciences*, 11(6), 2021.
7. Jan Konecny. On attribute reduction in concept lattices: Methods based on discernibility matrix are outperformed by basic clarification and reduction. *Inf. Sci.*, 415:199–212, 2017.
8. Utku Kose, Deepak Gupta, Victor de Albuquerque, and Ashish Khanna, editors. *Data Science for COVID-19 Volume 1. Computational Perspectives*. Academic Press, 2021.
9. S. Kuznetsov. On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49:101–115, 2007.
10. Sergei O. Kuznetsov and Artem Revenko. Interactive error correction in implicative theories. *International Journal of Approximate Reasoning*, 63:89–100, 2015.
11. Léonard Kwuida, Rokia Missaoui, Abdélilah Balamane, and Jean Vaillancourt. Generalized pattern extraction from concept lattices. *Annals of Mathematics and Artificial Intelligence*, 72(1):151–168, 2014.
12. Rokia Missaoui, Lhouari Nourine, and Yoan Renaud. Computing implications with negation from a formal context. *Fundamenta Informaticae*, 115(4):357–375, December 2012.
13. Angel Mora, Pablo Cordero, Manuel Enciso, Inmaculada Fortes, and Gabriel Aguilera. Closure via functional dependence simplification. *International Journal of Computational Mathematics*, 89(4):510–526, 2012.
14. O Postavaru, SR Anton, and A Toma. Covid-19 pandemic and chaos theory. *Mathematics and computers in simulation*, 181:138–149, 2021.
15. Massimo Pulejo and Pablo Querubín. Electoral concerns reduce restrictive measures during the covid-19 pandemic. *Journal of Public Economics*, 198:104387, 2021.
16. Jan Triska and Vilém Vychodil. Logic of temporal attribute implications. *Ann. Math. Artif. Intell.*, 79(4):307–335, 2017.