The data analytics of Covid-19 (or any other virus for that purpose)

João Pires da Cruz
Co-Founder & Partner *
(Dated: March 13, 2020)

The emergence of the Covid-19 pandemy and the different national reactions to the spreading of the disease showed that there is still a lot to learn about epydemics, while also poses a big challenge on those of us that have to deal and understand data and retrieve information from it. This paper is about how we, at Closer, have been discussing the issue and how we are understanding the stream of data that such a crisis produces. An important warning must be made **this is not a medical paper and does not, by any means, intended to contradict, substitute or discuss actions from medical and health authorities**

I. INTRODUCTION

Putting science before data in data science has been one of our methodologic advantages in the market. In fact, it is a very rare event that a set of data reflects the entire set of mechanisms that produces that data. Usually, the set of data that we collect from a non controled experiment, like all social events are, depends much strongly on the conditions in which the data was collected and on the nature of the mechanisms that produce it. That is our objective when we ask 'What is the physics of the problem?'.

The Covid-19 pandemy was the first to happen in the midst of a global usage of electronic social networks that generates instantaneous spreading of data and misleading information. All the while, the democratization of access to information about what was happening in several countries at the same time promote the discussion arround a couple of metrics on spreading and severety of the disease never before seen. Things that puzzles even the more educated people that challenged authorities, revealed that authorities were not used to comunicate a public health problem to the common citizen and showed medical authorities with similar credibility assuming completly contraditory positions.

This is naturaly a very complex situation and our mission is *to challenge complexity*. This paper is about how we would build data analytics arround these phenomena. We do not intend, by any means, to communicate a medical experiment or to analyse and/or criticize the actions of any of the national authorities involved. Our intention is only to provide the reader an example of rational reading of data, beggining with the physics of the problem and the mathematical challenges it brings. Based on that we will try to understand how different national authorities took different decisions based on exactly the same data, which is by itself a very interesting issue in decision support design, but also try to understant theoretically why countries had so different outcomes.

The goal is not to give an extensive description of what happened, that we will leave to medical researchers which are much more qualified for that. The goal is to theoretically understand the mechanisms before looking at data and then try to understand where data does not fit the theory, i.e., why is data different from what should be expected from logic and how it can promote different visions about how to tackle the epidemic problem. From there we will try to understand how such different perceptions can appear and finally, we will try to design a decision support metric to understand the evolution of the epidemy. In the end, this is not about epidemics, it is just about physics, the data that physics produces and how a mismatch between logic and data can promote decision diferences.

II. ELEMENTARY MY DEAR WATSON

Back in the 19th century in England, the tracing of aristocratic names was a big issue. That was what brought Francis Galton, a statiscian from that era, to ask the question about how probable it is for an aristocratic name to be extinguished? Bear in mind that the name would pass from father to son, so the female offspring would represent a lower probability of spreading of the name. The question was answered by Rev. Henry Watson that published (see Fig 2) what become known as the Galton-Watson process[1].

On the PROBABILITY *of the* EXTINCTION *of* FAMILIES. By the Rev. H. W. WATSON. With PREFATORY REMARKS, by FRANCIS GALTON, F.R.S.

FIG. 1. Watson paper

The Galton-Watson process become the first of a class of stochastic processes known as *branching processes*[2]. Let us think of objects that can produce aditional objects of the same kind. In each instant, the existing objects produces new objects that will produce new objects in the next instant and so on. We will call Z_n the random variable that represents the number of such objects in the n-th generation and assume that $Z_0 = 1$.

Let $P(Z_1 = k) = p_k$ the probability that an object in the 0-th generation produces k children and let us assume that this probability does not change from generation to

* joao.cruz@closer.pt

generation. This is the only parameter of the problem that can be taken as statistically measurable for now, because it does not depend on history. With this we can also define $P(i,j) = P(Z_{n+1} = i|P(Z_n) = j)$, that is, the probability of the generation $n+1$ has i objects knowing that the generation n has j objects, which obviously only depends on p_k that we assumed as constant from the first until the last generation. We wil call $P(i,j)$ as the transition probability.

We can go a bit deep on the mathematics, by defining f as the probability generating function of p_k, i.e.,

$$f(s) = \sum_{k=0}^{\infty} p_k s^k \qquad |s| \leq 1 \tag{1}$$

The Eq.(1) can be expressed in terms of transition probabilities

$$f(s) = \sum_{k=0}^{\infty} P(k,1) s^k \qquad |s| \leq 1 \tag{2}$$

where $P(k,1)$ is the probability of having k objects in generation 1 knowing that generation 0 has 1 element, exactly the same as p_k. We can write generically that

$$[f(s)]^i = \sum_{k=0}^{\infty} P(k,i) s^k \tag{3}$$

which means that whatever we have in the 0-th generation, the process for each object will agregate by multiplication (e.g., if 1 generates 3, 3 generates 9).

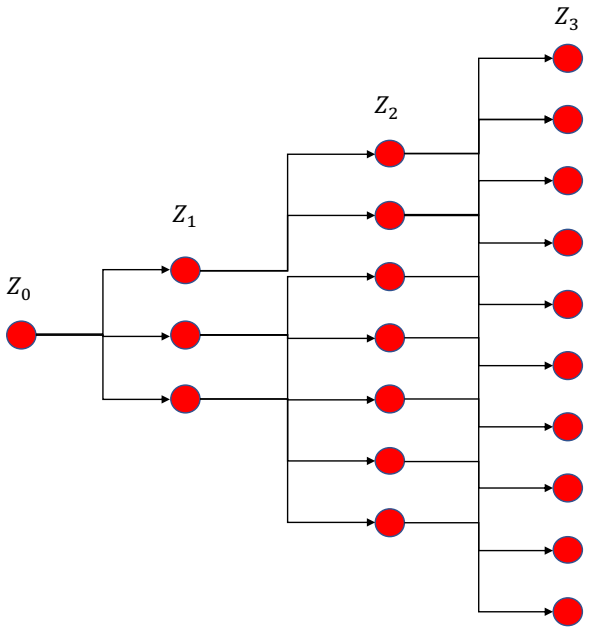

FIG. 2. Generation production

Since this generation producion is the real statistical event (because the system is inflating, it is not stationary) we must define $f_0(s) = 1$ and Eq.(1) will give $f_1(s) = f(s)$. So, taking $f_{n+1}(s)$ as the probability generating function of Z_{n+1}, under the condition that the realization of $Z_n = k$ has a distribution given by

$$f_{n+1}(s) = \sum_{k=0}^{\infty} p_k [f_n(s)]^k = f_n[f(s)] \qquad |s| \leq 1 \tag{4}$$

and this leads us to get the moments of Z_n as the derivatives of f at $s = 1$[3]. So if we call m to the expected moment of Z_1, $E[Z_1] = f'(1) = m$[4]. Applying the same rule,

$$E[Z_n] = f_n'(1) = [f[f_{n-1}(1)]]' = f_{n-1}'(1) f'(1) = \cdots = [f'(1)]^n = m^n \tag{5}$$

The same reasoning can be applied to obtain the variance of Z_n with the second derivative (which we will not show, see references) to get

$$Var(Z_n) = \begin{cases} \frac{\sigma^2 m^{n-1}(m^n - 1)}{m-1}, & \text{if } m \neq 1 \\ n\sigma^2, & \text{if } n = 1 \end{cases} \tag{6}$$

where $\sigma^2 = Var[Z_1]$.

This means that if we can assume that the distribution of Z_1 has a finite variance, so the Central Limit Theorem applies and we can unserstand the entire process just knowing m and σ, which as parameters of Z_1, and n which is the generation.

Naturaly we can take the units we want, so whatever we called 1 in Z_0 can be an arbitrary unit and, also, we can make the approximation to continuous generation by assuming that n is the time.

Also, if we call $x(t) = E(Z(t)) \approx E(Z_n)$ we know that $x(t) = m^t$ and $x(t-1) = m^{t-1}$, so

$$\frac{x(t) - x(t-1)}{x(t-1)} = \frac{m^t - m^{t-1}}{m^{t-1}} = m - 1 \tag{7}$$

and this means

$$x(t) = x(0) e^{(m-1)t} \tag{8}$$

This gives us a way to characterize the evolution of generations just by fine tunning m. So if $m > 1$ the process is explosive, the number of objects grow exponentially and we call it supercritical. If $m < 1$ the process is a decay, as time goes by there will be less and less objects until it disapears and we call it subcritical. If $m = 1$ the number of objects does not change *on average* with time, it keeps fluctuating between a supercritical regime and a subcritical regime like in a phase transition and we call it critical.

If we apply it to the aristocratic names, having a son favours m, having daughters lowers it. The take home messages from this section is that at any generation we can know the expected number of objects by knowing the

distribution Z_1 and m, the expected value of Z_1 defines the entire behavior of the process.

But what happens in practice? Unless we have a complete knowledge about the mechanisms of generation process that leads us to a rigourous knowledge of p_k, the only way to get m and σ^2 is to use statistical tools to measure every reproduction event, meaning taking every object and measure how many descedents it has and build the distribution.

Now, instead of a sexual reproduction we will assume that the reproduction is assexual and we are not interested on the propagation of the family name but on the object it self. If all offspring is viable, measuring all the reprodution events will give the average fertility. Now imagine that some offspring are not viable and die (see Fig. 3).

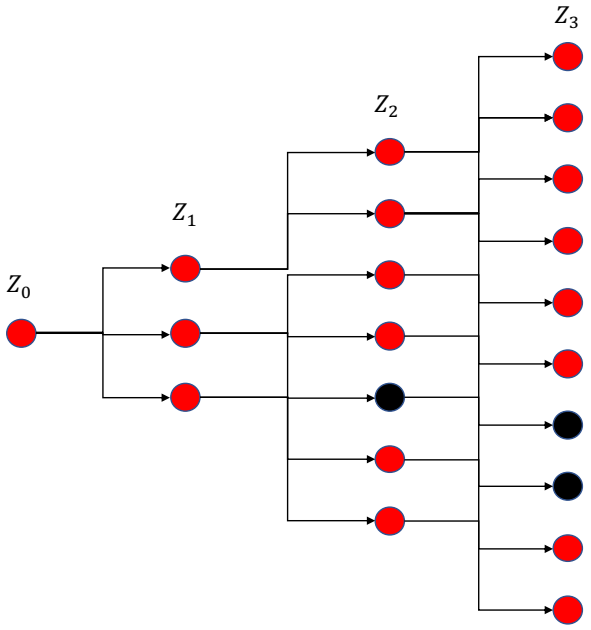

FIG. 3. Generation production with mortality

When we measure the average fertility, death will influenciate the measure of p_k and, in the end, of m and σ^2. Moreover, the influence of birth is the same as the influence of death in the measure of m. If the deaths are larger than births, the process will be subcritical and will end exponentially fast. If not, if deaths are less than births, then the process will be supercritical and the process will explode exponentially fast. If, on average, deaths are equal to births, then the process will be critical and will be 'fluctuating' arround constant.

In other words, since we usually do not know the distribution of Z_1 in advance, we have to measure it from the data and whatever measure we make, it will be influenced by deaths and births.

Another thing we did not mentioned yet is the existence of an universe in which the objects propagate. If we think of a virus (finally), it does not reproduce in the air, it needs a host (see Fig. 4). Ignoring deaths for now, as the reproduction process is far from the limits of the system, meaning the number of contaminated hosts are much smaller than the overall number of hosts, the process behaves as if it is in a free environment and the measure of the distribution of Z_1 is equivalent to measuring it in an infinite size environment.

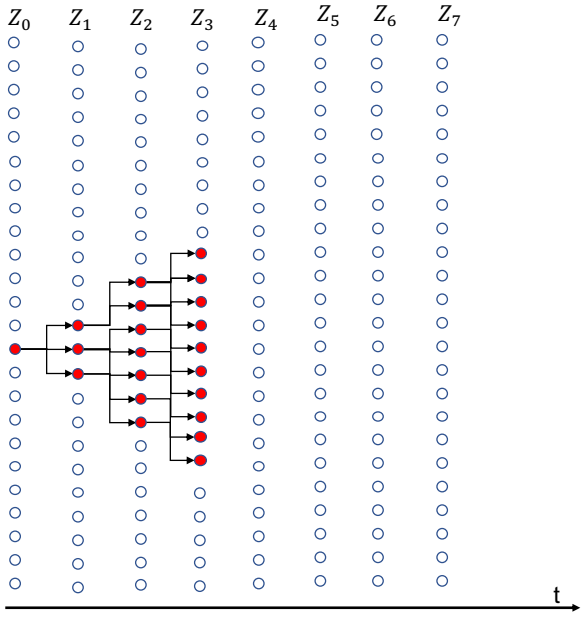

FIG. 4. Generation production with large reservoir of free hosts (white non-contaminated hosts, red contaminated hosts)

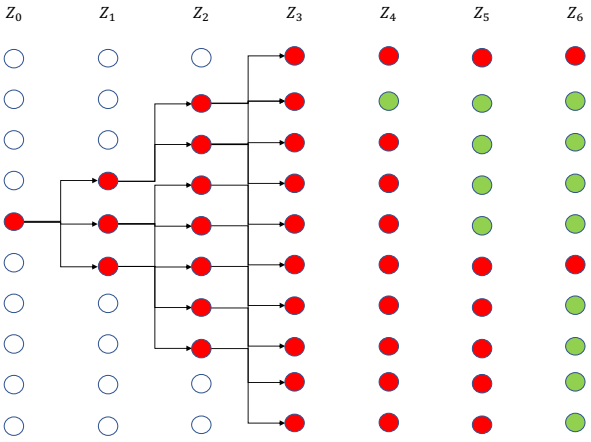

FIG. 5. Generation production with large reservoir of free hosts (white non-contaminated hosts, red contaminated hosts, green cured hosts)

But in the real world there is no such thing as an infinite environment. If we are speaking about a universe composed by hosts, then the number is finite. As the process evolves, the number of contaminated hosts begins to be in the order of the size of the system (see Fig. 5). When this happens the number of possible hosts to be

contaminated starts to shrink as the number of contaminated ones grows until it is not possible for the reproduction process to expand more. Statistically, the measure of m will start to approach 1, independently from what we measure when the reprodution was free. So as the process experiences the boundaries of the universe of hosts, it becomes critical.

If the hosts can be contaminated again (or if the hosts are substituted) then the process remains critical. If not, and as the hosts become cured, for the reprodution process is like a shrinking of the universe and the realizations of Z_n become lower than Z_{n-1}. The process becomes subcritical because the measure of m will be less than 1.

These very simple mathematics, models the transmission of viral agents, both in living beings, in internet sites, in nuclear plants or in economic environments.

Take the example of nuclear fission. When a neutron hits a nucleous, it excites it and it brakes in several pieces and produces between 3 and 5 new neutrons that will hit a new nucleous and so on. In this example the host is the nucleous and the virus is a neutron. What nuclear engineers do is to create a big concentration of nucleous in order that the universe of hosts become higher and this creates what is known as chain reaction, i.e., the nuclear process becomes supercritical. In order to avoid the explosion of the reactor, nuclear plants use techniques like putting carbon bars in the middle of the nuclear cells to lower the number of neutrons produced in each fission. When they do this, the process becomes subcritical. So the good operation of the nuclear plant is to alternate between supercritical and subcritical regimes to turn the process critical.

Obviously, when the nucleous is broken it is like a cure and it cannot be used again, so the system shrinks to the point it becomes subcritical and a replacement of the cells is needed.

In the next section we will be dealing with our goal which is to read the Covid-19 numbers, reactions and numbers from a data analytics perspective.

III. IS THERE A DOCTOR IN THE HOUSE?

Now, let us go back to our initial problem which is not about aristocratic names, but about epidemics. Again, we emphasize that this is just a data analytics set of thoughts and **not a medical paper**. Is just physics and mathematics.

Remember that both Italian and Spanish authorities decide not to react imediatly, in opposition to Portuguese authorities. Several medical doctors from those countries said that the Covid-19 epidemics was not more serious than seasonal flu. Based on World Health Organization in its February report about Covid-19[5], several news papers anounced that Covid-19 is more contagious and more deadly than the seasonal flu virus. Meaning that based on the exact same data, experts from different countries toke different decisions

Mathematically, and using the reasoning from last section, the statment that it is more contagious and more deadly is not possible because, as we saw, being more deadly contribute negatively to m and a lower m means less contagious. If it is more contagious, it means a bigger m and less deadly. So mathematically speaking the Italian and Spanish authorities would be theoretically right.

But, as data scientists we know two things:

1. Virus do not like Coreans more than Chinese, or Italians more than Germans;
2. Medical doctors deserve our respect, both the ones that decided not to act and those who decided otherwise.

So, can there be a data analytics explanation for the differences in opinion and for the differences in the numbers between countries?

First, we are not completly aware on how the results are obtained. We know the theoretical basis, but can we use the numerical results that were colected directly into our model? The difference between countries show that no, we cannot. The way data is colected is conditioned to whatever are the local health conditions, i.e., the way the data to get m is obtained, independently from the virus we are speaking. Obviously, medical doctors can have different ways to measure the hability for a virus to spread, namely from direct observations of the mechanism.

These differences can be qualitatively seen in the relative numbers, due to point number 1., virus do not like Italians more than Germans, i.e., it is not a physical diference, is a context diference, i.e., is how data is collected. An interesting project from a data science point of view is to study with medical authorities how to compensate the context to retrieve data that is compatible with the physics of the process.

Second, knowing the physics of the problem, we have an expectation about how the epidemy should behave. If we take Eq.(8) we know that

$$\frac{dx(t)}{dt} = x(0)(m-1)e^{(m-1)t} \tag{9}$$

meaning that if a process is supercritical, then its time derivative is also exponential. This is an important information if we want to evaluate the actions of public policies regarding an epidemy.

See Figs. 6,7 and 8. They represent the evolution of daily confirmed cases and suspects of Covid-19 for two and half months for three different countries, Portugal, Corea and the UK. Both Portugal and Corea are not in a supercritical process. Whatever the measures taken by local authorities, they seem to be working and, in the case of Corea it is clearly in a subcritical phase, while Portugal seems to be following a critical process.

Now let us compare with the UK. Seeing the UK, both in cumulative numbers and the derivative, they seem both to be exponentials meaning that the UK is in a

supercritical process, in other words, the UK needs to do more to takle the epidemy (assuming that they want to).

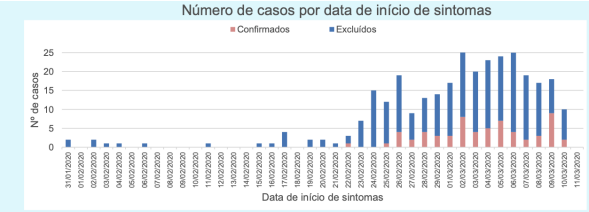

FIG. 6. Portugal evolution of new cases of Covid-19 as-of March 12th, 2020 (source: Portugal Health Ministry)

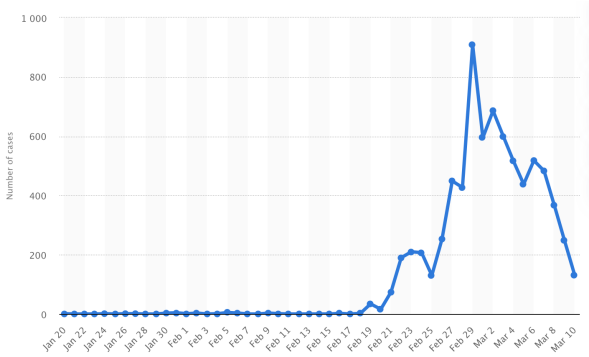

FIG. 7. Corea evolution of new cases of Covid-19 as-of March 12th, 2020 (source: Statista)

Based on our discusion on branching processes and our example of the nuclear plant, we can guess that both Portuguese and Corean authorities took measures to turn the process subcritical or, at least, critical while the UK is not dropping their 'carbon bars' in the nuclear fuel cells.

IV. CONCLUSIONS

This small exercise intends to show to data scientists the importance of finding the answer to the question 'what is the physics of the question?' to understand the biggest problem of the Covid-19 crisis data science wise. Why countries produce so different results and why the same data can lead to different decisions in highly respected physicians? We can speculate that the difference between theory and practice can help on the answers to these questions. And that difference is amplified by data that is produced in special context without compensation. Also we can use theory to see that some country policies are working much better than others.

But the main conclusion is that it does not matter the amount of data you collect if you don't know the mechanisms that produce the data and if you do not know how to hadle theory to find the answers that data can give you.

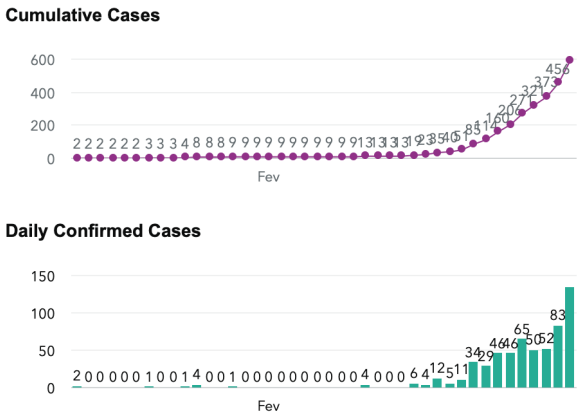

FIG. 8. UK evolution of new cases and cumulative of Covid-19 as-of March 12th, 2020 (source: Public Health England)

ACKNOWLEDGMENTS

The author would like to thank to Dr. Raul Vaz Osório for checking for misleading medical information and João Leal Fernandes for revision.

V. BIBLIOGRAPHY

[1] H. W. Watson and F. Galton, On the probability of the extinction of families, The Journal of the Anthropological Institute of Great Britain and Ireland **4**, 138 (1875).
[2] T. E. Harris, The theory of branching process, (1964).
[3] K. Athreya, P. Ney, and P. Ney, *Branching Processes*, Dover Books on Mathematics (Dover Publications, 2004).
[4] N. Johnson, A. Kemp, and S. Kotz, *Univariate Discrete Distributions*, Wiley Series in Probability and Statistics (Wiley, 2005).
[5] *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*, Tech. Rep. (World Health Organization, 2020).