RELATION OF COVID-19 INCIDENCE AMONG BRAZILIAN REGIONS USING STATIC AND DYNAMIC BAYESIAN NETWORK.

Mariana Thais Almeida[1], Vinícyus Araújo Brasil[2], Lilia Carolina Carneiro da Costa[1], Anderson Ara[2]

[1] Federal University of Bahia, Salvador, Brazil

[2] Federal University of Paraná, Curitiba, Brazil

Abbreviated abstract

This work consists of studying the relationship of the dynamics of the pandemic of COVID-19 among Brazilian regions, in order to assess whether the increase in the incidence of the disease in a given region implies the worsening of the pandemic in another region. Thus, an analysis of the relationship of the incidence rates of COVID-19 among the regions of Brazil is proposed by joining the methodologies of time series and static Bayesian networks, and comparing them with dynamic Bayesian networks. The use of these two methodologies together is presented in some works such as [3].

Email: mariana.thais@ufba.br

METHODOLOGY

Database

The analysis period was from 02/25/2020 to 02/15/2022, and for each region the **number of new cases per 100,000 inhabitants in each of the 104 epidemiological weeks** was considered.

Filtering the series via ARIMA models

The model can be described mathematically as:

$$(1 - \phi B) \bigtriangledown^d X(t) = (1 - \theta B)\epsilon(t),$$

where ø(B) is the autoregressive operator of order p, △ X(t) is the differentiated series, Θ(B) is the autoregressive moving average operator of order q and ε(t) is white noise.

Bayesian Network

Bayesian network (BN) consist of a visual and informative **representation of the joint distribution of all variables involved in a problem**.

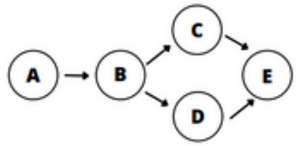

Figure 1: Example of a Bayesian Network.

Let Y be a continuous random variable and X_i, i = 1,2,...,k a set of continuous parent variables of Y, we say that Y is a **Gaussian linear model** if there are the parameters β_0, β_1, ..., β_k and σ² such that

$$p(Y|\mathbf{x}) = \mathcal{N}(\beta_0 + \beta^T \mathbf{x}, \sigma^2).$$

Static RB does not allow the use of time observations, and an extension of this approach that represents graphical models for stochastic processes is the **dynamic Bayesian network.**

The nodes are observed sequentially at discrete times represented by Y_i(t), which is defined as a process i performed at time t, and the arcs represent the evolution from one time to another. For the adjustment and estimation of dynamic Bayesian network parameters some assumptions are adopted, such as **first order Markov** and **time homogeneity**.

Structure Learning

Network learning **is performed in two steps [4], the estimation of the network structure and the estimation of the parameters.**

The static network structure learning algorithms were **PC-Stable, Hill Climbing (HC) and Tabu [2]** and the parameter estimates were obtained through the maximum likelihood estimator. To find possible directions for the arcs we used **bootstrap sampling and average model proposed by Friedman [1].**

Dynamic Max Min Hill Climbing (DMMHC) [5] was used for dynamic structure learning.

RESULTS

For each of the series an ARIMA(p,d,q) model was fitted, and the best model for each region was found using Akaike's Information Criterion (AIC).

Table 1: Estimated coefficients and their standard error.

	Norte	Nordeste	Centro-Oeste	Sudeste	Sul
ϕ_1	1,3974	0,0758	-	-	-
	(0,1612)	(0,1065)	-	-	-
ϕ_2	-0,6887	0,2802	-	-	-
	(0,0902)	(0,1064)	-	-	-
θ_1	-0,8011	-	-0,1250	-0,2862	-0,2336
	(0,1726)	-	(0,0936)	(0,0922)	(0,1054)
θ_2	-	-	0,3507	-	-
	-	-	(0,1026)	-	-

From the residuals of each of the adjusted univariate models, the structure was adjusted and the parameters of the static RB were estimated.

The **1000 simulations were performed for the PC, HC and Tabu algorithms** and for all possible combinations of arcs two by two it was observed the frequency with which they appear in the learned networks and the probabilities of the directions of each arc conditional to the arc being present in the graph.

Then **the arcs with a greater than 50% probability of being present in the network and the most plausible direction were selected.**

For inference, the models **were compared using the BGe** (Bayesian Gaussian equivalent) score and **the Tabu model was selected.**

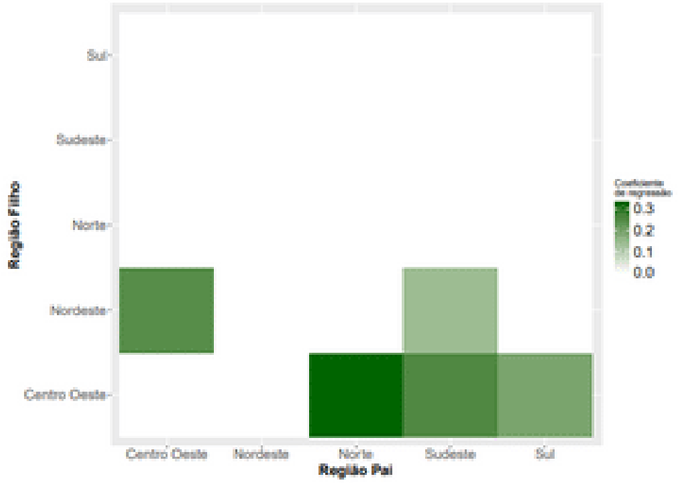

Figure 2(a): Regression coefficients of the model.

In the heat map, the value of the coefficients represents the strength of the arc and the darker the color in the graph the stronger the relationship of impact of one region on the other in contamination.

It can be seen that all the directed relationships are positive, so that, considering the other fixed variables, **the rate of COVID-19 in the son region increases as there is an increase in the rate of new cases of the disease in the parent region.**

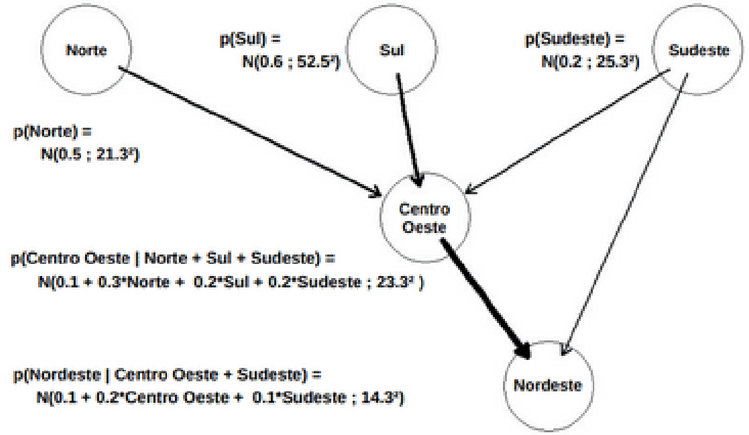

Figure 2(b): Marginal node distributions.

In Figure 2(b), the thicknesses of the arcs represent the frequency with which these arcs appear in the structure estimated by the Tabu model and the conditional distributions of each region are shown.

Figure 3 shows the estimated structure where the nodes ending _t_0 represent the regions in the present and the nodes ending _t_1 represent the regions in the immediate past.

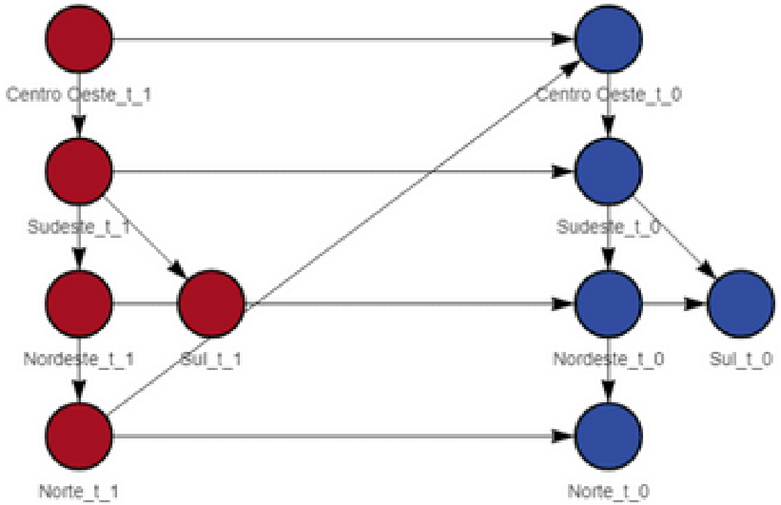

Figure 3: Estimated structure of the dynamic Bayesian network.

All regions are dependent on their immediate past and the only region at time _t_0 that has a direct relationship with another region's past is the Midwest (influenced by the North at time t-1).

Table 2: Conditional marginal distributions for each region over time.

	P(Região_t_1 | pais)
Região_t_1	
Centro-Oeste	$\mathcal{N}(162,36;90,81^2)$
Nordeste	$\mathcal{N}(9,94+0,80Sudeste\_t\_1;34,38^2)$
Norte	$\mathcal{N}(33,40+0,84Nordeste\_t\_1;42,64^2)$
Sudeste	$\mathcal{N}(11,41+0,62CentroOeste\_t\_1;35,84^2)$
Sul	$\mathcal{N}(5,79+1,40Sudeste\_t\_1;72,64^2)$
Região_t_0	
Centro-Oeste	$\mathcal{N}(5,56+0,85CentroOeste\_t\_1+0,17Norte\_t\_1;26,96^2)$
Nordeste	$\mathcal{N}(2,84+0,88Nordeste\_t\_1+0.09Sudeste\_t\_0;17,51^2)$
Norte	$\mathcal{N}(5,90+0,92Norte\_t\_1+0,03Nordeste\_t\_0;20,55^2)$
Sudeste	$\mathcal{N}(3,65+0,67Sudeste\_t\_1+0,21CentroOeste\_t\_0;25,73^2)$
Sul	$\mathcal{N}(6,09+0,69Sul\_t\_1+0,40Sudeste\_t\_0;53,98^2)$

For all regions there is a decrease in variability over time. At time _t_1, the region with the greatest variability is the Midwest, followed by the South. At time _t_0 the South has the highest variability, a result also observed in the static network. Moreover, the strength of the parent regions tends to decrease from one time to another.

DISCUSSION

The Southeast region is pointed out as the greatest influencer in the rates of new cases of the disease, because in the static approach it is the region that directly influences the largest number of regions and in the dynamic approach it has the most connections.

The Midwest is also one of the most influential regions in the dissemination of COVID-19, since it presents more connections in the static Bayesian network and in the dynamic network it influences the Southeast. In both approaches the North influences the Midwest and the Northeast is influenced by the Southeast. Moreover, when the immediate past of a region is known, the strength of the other regions influencing it decreases.

REFERENCES

[1] FRIEDMAN, N.; GOLDSZMIDT, M.; WYNER, A. J. Data analysis with bayesian networks: A bootstrap approach. In: UAI. [S.l.: s.n.], 1999.

[2] GLOVER, F.; LAGUNA, M. Tabu search. Kluwer Academic Publishers, v. 3, p. 621–757, 1998.

[3] QIU, J. et al. Spatial transmission network construction of influenza-like illness using dynamic bayesian network and vector autoregressive moving average model. BMC Infectious Diseases, 2021.

[4] SCUTARI, M.; DENIS, J.B. Bayesian Networks With Examples in R. [S.l.]: CRC Press, 2015.

[5] TRABELSI, G. New structure learning algorithms and evaluation methods for large dynamic Bayesian networks, 2013.