Introdução a Ciências de Dados

Jorge Augusto Salgado Salhani

I. Introdução

A análise do comportamento de sistemas reais é um tópico muito abordado recentemente em sua grande parte devido ao aumento sem precedentes da quantidade de dados coletados e armazenados [4]. A atenção dada a este cenário permite a construção de pesquisas e aparatos de inovação de forma muito mais recorrente e abrangente, o que garante sua utilização em ramos multidisciplinares em ambientes médicos e de saúde [11], de cibersegurança [2, 16] até na análise de sinais [24] (aplicado, por exemplo, em sistemas de comunicação).

O conteúdo presente na área de ciência de dados é extenso. Análise preditiva, descritiva e profunda representam alguns deles [4], os quais são construídos de forma sólida sobre métodos matemáticos advindos de sistemas dinâmicos, modelos estocásticos, análise complexa, estatística, entre outros.

Neste sentido, o conteúdo apresentado neste trabalho busca sumarizar partes importantes do que foi ensinado na aula de Introdução a Ciência de Dados (SCC0275), conduzida remotamente pelo prof. dr. Rodrigo Fernandes de Mello (ICMC - USP), além de alguns detalhamentos mais teóricos por meio de pesquisas na literatura (selecionadas pessoalmente ou sugeridas ao longo do curso). As aulas podem ser acompanhadas abertamente na plataforma *YouTube*, no canal de nome **ML4U**[1].

Por conta do extenso conteúdo lecionado, optamos por separá-lo em duas grandes estruturas de análise: 1. Aprendizado Baseado em Instâncias; 2. Análise de Séries Temporais. Desta forma, este trabalho contém 5 seções denominadas **Motivação de Estudo**, **Fonte de Dados**, **Metodologia**, **Resultados** e **Conclusão**, as quais serão subdivididas em duas sub-seções (excluso a Conclusão) relativas ao problema escolhido que será tratado por meio do Aprendizado Baseado em Instâncias (*A.* IBL) (acrônimo do inglês **Instance-Based Learning**) e ao problema abordado via Análise de Séries Temporais (*B.* Séries Temporais).

As análises serão realizadas sobre problemas políticos relacionado à governanças *A. IBL* e sobre flutuações do preço de medicamentos *B. Séries Temporais*. Deste modo, na seção *Motivação de Estudo* apresentaremos uma visão geral sobre cada problema escolhido e qual sua importância na vida das pessoas; na seção *Fonte de Dados* mostraremos quais foram as bases de dados escolhidos sobre as quais realizaremos as análises propostas; em sequência são apresentadas as bases teóricas utilizadas por cada modelo na seção *Metodologia*; por fim, mostraremos os resultados obtidos para cada cenário (*A.* e *B.*) na seção *Resultados*.

Nas seções finais de *Conclusão* e *Referências* encontram-se o que pudemos concluir deste trabalho e a bibliografia utilizada sob a qual embasamos este estudo.

[1]Disponível em: https://www.youtube.com/c/ML4U_Mello/featured

II. Motivação de estudos

A. IBL

O estilo de governo de uma determinada região apresenta relação direta com a forma de vida levada pelas pessoas que vivem sob ele. Por exemplo sistemas democráticos são reconhecidos por estarem vinculados a melhores condições de educação [6] e sentimentos de paz e segurança [22, 3].

Em oposição ao que mencionamos a respeito de democracias, governos autocráticos restringem a participação política de cidadãos, a manutenção de poder é garantida por regras de sucessão e chefes executivos são dotados de poder cuja verificação por poderes repartidos (legislativo, executivo e judiciário [19], por exemplo) é parcial ou completamente omissa [13] o que acompanha políticas de direitos desiguais. Regimes autocráticos e democráticos são vistos como contrastantes por conta dos meios como poderes são distribuídos, como a ordem social é realizada e o modo como políticas de cooperação e regulação são estabelecidas [13] entre outros fatores.

Desta forma, é importante que seja possível estabelecer métricas que possibilitem caracterizar um regime político conforme atributos quantificáveis. A pontuação POLITY [13] é um recurso particularmente interessante que utilizaremos em nossa análise, o qual atribui valores inteiros no intervalo $[-10, +10]$ capazes de associar governos institucionalmente autocráticos como -10 e governos institucionalmente democráticos como $+10$. Esta métrica será utilizada como rótulo de classificação em nosso modelo utilizando IBL.

Por outro lado, como esta pontuação considera qualidades práticas de instituições políticas tais como ações de poderes executivos e competição política (e.g. poderes de oposição institucionalizados), nossa aplicação do modelo de classificação busca relacionar métricas econômicas (PIB, média de horas trabalhadas pela população, etc.) como potenciais preditores do índice político em um determinado ano.

B. Séries temporais

Grande parte dos sistemas reais (se não todos!) apresentam evolução temporal. As áreas sobre as quais análises de séries temporais podem ser aplicadas são incontáveis. Podemos citar em métricas econômicas [10, 8, 11], relacionadas à transmissão de doenças [21] até a medidas da taxa de batimentos cardíacos [31].

Neste sentido, a avaliação de uma série de forma a predizer seu comportamento para tempos futuros baseado no conhecimento da série em eventos passados é importante. A predição de acidentes, controle de doenças e predição de desastres [25] são alguns dos contextos que motivam nosso estudo sobre este tema.

A disponibilidade de medicamentos de fácil acesso para a população é um fator importante para melhorar a qualidade de vida das pessoas tanto no nível de informação apresentada às pessoas quanto no que tange ao acesso financeiro a eles [11]. No Brasil temos, por exemplo, a implementação institucional do programa Farmácia Popular [1], promovendo redução de mortalidade e de internações por doenças crônicas.

Com isso, a predição de custos de dado medicamento (em nosso caso, medicamentos analgésicos [11]) pode representar um importante fator para pacientes que dependem de um composto, especialmente caso seu costo seja muito elevado e/ou com baixa disponibilidade no mercado.

III. Fonte de dados

A. IBL

Para a análise proposta, serão considerados os dados disponíveis em acesso aberto compilados por por Max Roser e Hannah Ritchie, com excelente visualização no site *Our World in Data* [26]. A análise dos estilos de governo inclui dados de relações internacionais, [20] igualdade de gênero [17] e a compilação de dados sobre a democracia de cada país considerado, também feita por Max Roser. [27]

B. Séries temporais

A análise de séries temporais apresenta um conjunto de dados mais simplificado, já que temos um único observável (custo de medicamento, em dólares norte-americanos) medido em um período de aproximadamente 3 anos (2 de janeiro de 2011 a 30 de julho de 2014, totalizando 1306 observações) [11]. Os dados originais são disponibilizados por este mesmo artigo.

Com isto, chegamos no conteúdo central deste estudo e para que seja possível uma navegação mais fácil pelos conteúdos, separamos abaixo os tópicos que cobriremos para nossas análises. Para estudos de governança (IBL) utilizaremos dos modelos:

- PCA (*Principal Component Analysis*);
- KNN (*K-Nearest Neighbors*);
- DWNN (*Density-Weighted Nearest Neighbors*);

Já para estudos de custo de medicamentos (séries temporais) utilizaremos os modelos:

- TET (*Takens embedding Theorem*);
- ACF (*Auto correlation function*);
- AMI (*Average Mutual Information*);
- FNN (*False-Nearest Neighbors*);
- EMD (*Empirical Mode Decomposition*);
- IMF (*Intrinsic Mode Functions*);
- Congruência de Fase;
- ARIMA (*Auto-Recursive Integrated Moving-Average*);
- PACF (*Partial Auto-Correlation Function*);

IV. Metodologia

A. IBL

Nesta seção focaremos na apresentação das ferramentas necessárias para a obtenção dos resultados sobre o tema de estilo de governanças, por meio do aprendizado baseado em instâncias. Utilizaremos como exemplo a base de dados Iris[2].

O modelo de PCA baseia-se primariamente na redução de dimensionalidade de um espaço de parâmetros por meio da obtenção de correlações entre múltiplas variáveis. Comumente uma determinada medida possui dependência com diversos parâmetros. Sendo que o número destes representa a dimensão dos nossos dados, por exemplo uma medida cuja dependência permeia 5 grandezas é dita com dimensão 5. Quando comparamos duas ou mais variáveis cuja correlação é expressiva, podemos resumi-las a uma única grandeza capaz de carregar a informação de todas as demais. A isso damos o nome de redução de dimensionalidade.

Quantitativamente podemos utilizar a correlação de Pearson para este fim. Esta medida é dada pela equação

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y))]}{\sigma_X \sigma_Y}, \tag{1}$$

tal que $cov(X,Y) \equiv E[(X-\mu_X)(Y-\mu_Y))]$ é interpretado como a covariância entre duas matrizes X e Y, μ indica

[2]Já incluso nas linguagens Python e R, mas pode ser obtidas pelo link: https://archive.ics.uci.edu/ml/datasets/iris

a media dos valores de cada conjunto de dados X e Y, $E[H]$ indica o valor médio de uma dada função H. Neste sentido, a medida de covariância representa qual a relação que dois conjuntos de dados possuem. Devido à normalização pelo desvio padrão σ, $\rho_{X,Y} \in [-1, +1]$ é adimensional.

A medida de covariância também pode ser compreendida por meio da equação

$$cov(X, Y) = X \cdot Y, \tag{2}$$

dada pelo produto escalar entre duas matrizes. Isto fornece a projeção dos dados subre uma reta de regressão entre os eixos X e Y

Uma vez encontrada a matriz de covariância, podemos encontrar os autovalores e autovetores associados à ela. Como autovetores indicam quais os pontos no espaço de parâmetros que sofrem apenas uma alteração escalar de modo que

$$T(\mathbf{V}) : \mathrm{Re}^n \to \mathrm{Re}^n \tag{3}$$

$$T(\mathbf{V}) : \mathbf{AV} = \lambda \mathbf{V}, \tag{4}$$

onde V é o vetor em estudo, A, a matriz de transformação e a última igualdade apenas ocorre quando V é dado por um autovetor da matriz A e λ representa um autovalor associado a V.

Neste sentido, sabendo que $V = I_n V$ (I_n é a matriz identidade), vale a relação

$$AV - \lambda V = 0 \implies (A - \lambda I_n)V = 0, \tag{5}$$

a qual apenas ocorre quando $\det[A - \lambda I_n] = 0$.

De modo mais intuitivo, quando determinamos os autovetores da matriz de covariância de um conjunto de dados, descobrimos qual a base ortonormal que melhor os representa e seus autovalores, qual a variância dos dados sobre esta base.

Assim, podemos aplicar este método da forma que segue. A base de dados *Iris* classifica flores em três espécies distintas (setosa, versicolor, virginica) de acordo com as seguintes medidas: comprimento e largura das sépalas e pétalas, resultando em 4 colunas de dados e, assim, fazendo com que dados associados a esta base apresentem dimensionalidade 4.

Quando comparamos duas destas medidas, por exemplo o comprimento e largura das pétalas, obtemos a relação apresentada na figura 1.

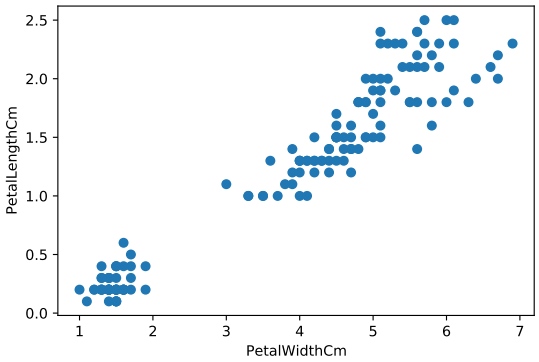


Figura 1: Relação entre comprimento e largura de pétalas para flores do gênero *Iris*.

Notamos que as duas grandezas na figura 1 possuem forte correlação positiva, o que significa que pétalas com largura crescente possuem, regularmente, pétalas com comprimento também crescente.

Como iremos comparar medidas que variam em intervalos não necessariamente igualmente espaçados, é importante que seja feita a normalização e centralização do espaço em que os dados são representados. Assim, mesmo aquelas que apresentem ordens de grandeza distintas têm contribuição próxima. Utilizando que $X_{escaled} = (X - E[X])/\sigma_X$ centraliza os dados ao redor da média $E[X]$ e escala de acordo com seu desvio σ_X, transladamos os pontos conforme visto na figura 2

Quando obtemos a matriz de covariância associada aos

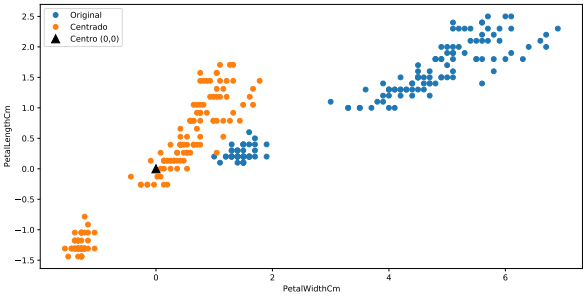


Figura 2: Relação entre comprimento e largura de pétalas para flores do gênero *Iris*. Representamos em conjunto os dados centralizados e escalados de acordo com o desvio padrão σ_X dos dados.

dados centrados e escalados, podemos obter os autovetores e autovalores por meio da equação 5. Assim, podemos representar os autovetores conforme a figura 3. Nela notamos que os autovetores representam uma base ortonormal rotacionada de modo que cada eixo esteja sobre a maior e menor variância do conjunto de dados. Além disso cada autovetor V_i ($i = 1, 2$, devido às duas dimensões) está associado a um autovalor λ_i que expressa a variância que cada autovetor representa. Em números, $\lambda_1 = 0.981$, $\lambda_2 = 0.019$. O primeiro está relacionado ao eixo com maior variância e sabemos que 98.1% da variância dos dados é dado por ela.

Com esta análise podemos concluir que dentre as duas variáveis comprimento e largura de pétalas, podemos utilizar apenas a projeção destes dados sobre o eixo correspondente ao autovetor V_1. Ao fazer isso, reduzimos para 1 as duas dimensões e mantemos 98.1% da informação contida nelas.

Uma das aplicações acontece pelo uso desta redução de dimensionalidade associado à categorização de dados. Quando desejamos obter classificação discreta de objetos, um dos modelos que estudaremos é o **KNN** (*K-Nearest Neighbors*, ou K-Vizinhos Próximos) e o **DWNN**

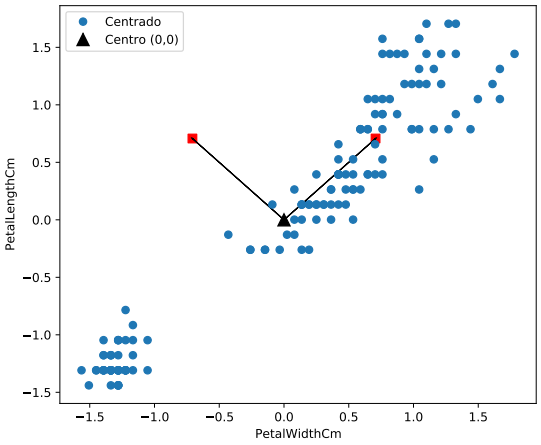


Figura 3: Relação entre comprimento e largura de pétalas para flores do gênero *Iris*. As retas em cor preta indicam segmentos dos autovetores (com comprimento unitário) associados à matriz de covariância dos dados comprimento e largura de pétalas.

(*Densitiy-Weighted Nearest Neighbors*, ou Densidade Ponderada de Vizinhos Próximos).

O algoritmo **KNN** (que utilizaremos para nossa análise) atribui a cada ponto do espaço um classificador conforme a base de conhecimento sugere. Mostramos a separação decorrente de cada espécie na figura 4.

Desta forma, dado um ponto de consulta, o algoritmo considera k vizinhos mais próximos desta posição e nos retorna qual a categoria mais provável à qual ele faz parte por meio da equação

$$\hat{f}(x_q) = \frac{1}{k} \sum_{i=1}^{k} f(x_i), \tag{6}$$

onde $\hat{f}(x_q)$ indica o atributo médio associado ao ponto de consulta (*query*) x_q e $f(x_i)$ ($i = 1, 2, \ldots, k$) o atributo de cada um dos x_i pontos mais próximos de x_q.

Com isso, mapeamos todo o espaço conforme nossa base de conhecimento e, em fim, dado um ponto que desconhecemos sua categoria (uma nova flor do gênero *Iris* que

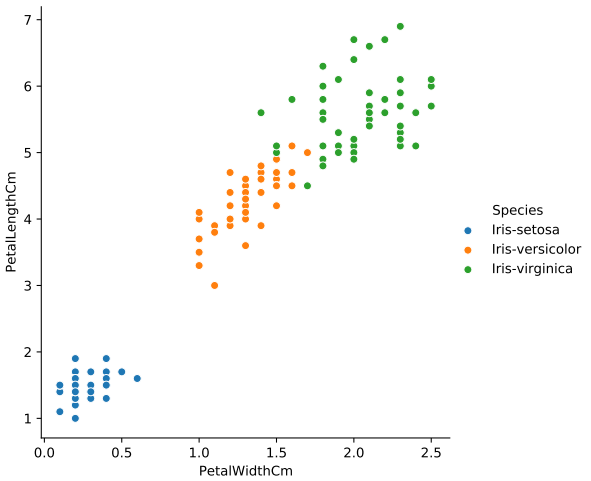


Figura 4: Relação entre comprimento e largura de pétalas para flores do gênero *Iris*, análogo à figura 1. As cores indicam cada uma das espécies presente na base de conhecimento.

tenhamos medido o tamanho de suas pétalas e sépalas, e.g.), podemos inferir qual sua espécie. Como utilizamos variáveis (ou instâncias) de um conjunto de dados (ou base de conhecimento) para inferir um determinado resultado ou seja, aprender com informações coletadas, temos a família de algoritmos denominada **IBL** (*Instance-Based Learning*, ou Aprendizado Baseado em Instâncias).

Um segundo modelo desta família de algoritmos que abordaremos a seguir é o **DWNN**. Sua concepção é análoga ao modelo **KNN**, porém mapeando um espectro contínuo, o que nos fornece a informação de uma densidade de probabilidade das categorias às quais que um ponto de consulta x_q pode estar presente. desta forma, temos a equação

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i}, \tag{7}$$

onde $\hat{f}(x_q)$ indica o atributo médio associado ao ponto de consulta x_q (análogo à equação 6). O parâmetro w_i é novidade nesta formulação, e está associado ao "peso"que determinado ponto x_i imputa à sua classificação $f(x_i)$.

Este peso w_i pode ser, a princípio, qualquer função. No entanto, uma bastante utilizada (e que também será útil para nosso estudo sobre regimes políticos) é dada por

$$w_i = \exp\left\{\frac{E[x_q, x_i]}{2\sigma^2}\right\}, \tag{8}$$

onde $E[x_q, x_i]$ nos retorna o valor médio entre o ponto de consulta x_q e o $i-$ésimo ponto da nossa base de dados e σ representa um parâmetro que controlamos e, na equação 8, indica a abertura da gaussiana centrada no ponto $E[x_q, x_i]$.

Tomando um vetor de densidade W tal que $w_i \in W$, temos que

$$\hat{f}(x_q) = \frac{W \cdot Y}{\sum_{i=1}^{N} w_i}, \tag{9}$$

onde obtemos o atributo médio $\hat{f}(x_q)$ do ponto de consulta através da soma ponderada (em forma de produto escalar) $W \cdot Y$, onde Y representa variáveis dependentes do vetor X e N, o total de pontos que consideramos de amostras pertencentes a X.

Por nos retornar um conjunto de valores reais, modelos desta natureza são baseados em **regressão**, o que difere de modelos por **classificação**, como é o caso do **KNN** que nos retorna valores enumeráveis no conjunto dos números naturais ou inteiros.

Um exemplo interessante e simples que podemos obter com este modelo é a regressão apresentada na figura 5. É possível notar que, quando consideramos aberturas de gaussiana pequenas ($\sigma = 0.1$), próximo aos pontos em azul (base de dados) retomamos uma forma de classificação discreta, uma vez que pontos mais distantes do que o primeiro encontrado terão contribuição negligenciável. Fato curioso ocorre exatamente no ponto médio entre dois pontos em azul. Como ambos apresentam o mesmo peso, qualquer deslocamento escolhe entre um deles, resultando em um degrau. Este fenômeno se dispersa (i.e., a regressão em vermelho torna-se mais contínua) conforme a abertura σ aumenta (caso $\sigma = 0.3$).

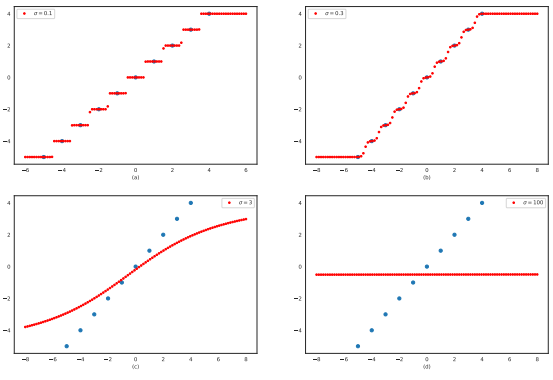


Figura 5: Regressão utilizando o modelo **DWNN** (pontos em vermelho) para um conjunto de dados (pontos em azul) linear. Consideramos, respectivamente, os valores de $\sigma = 0.1, 0.3, 3, 100$ para as figuras $(a), (b), (c), (d)$.

Para valores muito grandes, como $\sigma = 3$, a abertura é grande o suficiente para que a regressão apresente tendência à média dos pontos, com deslocamento apenas para valores extremos do nosso espaço de dados. Este efeito é destacado quando utilizamos $\sigma = 100$, e todos os pontos são considerados com mesmo peso e a regressão à média ocorre.

Em seguida voltaremos nossa atenção aos modelos da segunda parte da nossa análise, utilizadas para séries temporais.

B. Séries temporais

Apresentaremos agora os métodos que serão empregados para a análise de dados dispostos em formato de série temporal. Para que a análise final dos resultados apresentados na seção V-B possa ser compreendida em sua totalidade, utilizaremos sequências construídas manualmente (ou seja, nossas séries temporais consistirão de curvas senoidais com a presença -ou não- de ruído gaussiano).

Uma das análises possíveis sobre séries temporais ocorre através da predição (ou *forecast*) de seu comportamento ao longo do tempo. Para isto, precisamos definir métodos capazes de aprender com um espaço de eventos passados para estimar novos valores que não fazem parte do conjunto de dados original. Assim, nos deparamos com um problema de utilizar unicamente modelos regressivos, tal como o já mencionado **DWNN**.

Na figura 5, especialmente para $\sigma = 0.1, 0.3$, fica evidente que tal modelo prediz o comportamento de valores pertencentes à amostra, mas falha para os casos extremos, onde a regressão não prediz um comportamento linear com inclinação 1, mas constante (inclinação 0, na figura 5(d)).

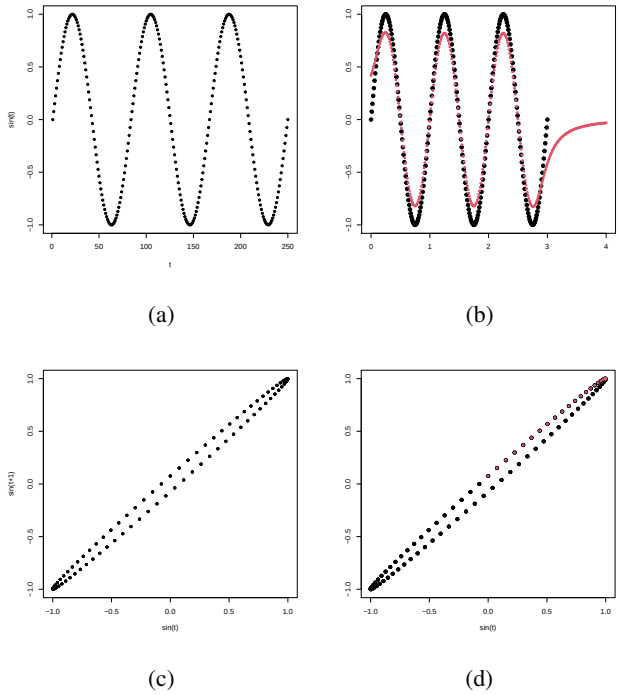
Uma abordagem que podemos utilizar é chamada *Taken's embedding theorem* (**TET**, ou Teorema de Inserção de Takens). Sua motivação vem da possibilidade de reorganizar o conjunto de dados na forma $x(t) \rightarrow F[x(t-d), x(t-2d), x(t-3d), \ldots, x(t-md)]$, de modo que uma medida no tempo t possa ser descrita como função de valores anteriores. Isso permite que um observável cuja dependência tenha sido primariamente medida em função do tempo possa ser descrito por meio de um número m deste mesmo observável medido em tempos anteriores.

Temos portanto dois novos parâmetros que serão importantes para a modelagem via *Takens' embedding* chamados de *embedding dimension* (ou dimensão inserida) que será referido como m, e *time-lag* (ou atraso temporal), referido como d. Neste novo espaço de m parâmetros com atraso d. Este novo espaço é denominado **espaço de fase**.

Quando no espaço de fase, os dados para prever o futuro são condizentes com os dados do passado, tornando um valor de teste $x(t)$ coerente com a uma predição para $x(t+1)$. Vale ressaltar que esta avaliação supõe que os dados sejam determinísticos.[5, 18]

A vantagem de representação deste novo espaço é clarificada quando comparamos os resultados de predição para

o modelo **DWNN** realizado sobre o espaço-tempo com os resultados com aprendizado sobre o espaço de fase, mostrados na figura 6. Os 50 pontos finais (em vermelho na figura 6(d) são preditos de modo muito preciso. O oposto ocorre para os pontos finais no espaço-tempo na figura 6(b).



(a) (b)

(c) (d)

Figura 6: Evolução temporal de curvas $\sin(2\pi t)$ (a) com pontos originais (preto) e preditos (vermelho, em (b) e (d)) utilizando o modelo **DWNN** com $\sigma = 0.1$. A simples representação em espaço de fase com um único atraso temporal (c) $x(t)$ *vs.* $x(t+1)$ faz com que o modelo preveja pontos de forma bastante acurada em (d).

Embora a análise ocorra corretamente neste modelo, sua aplicação não é imediata quando estudamos fenômenos reais. Estes eventos não são puramente determinísticos, o que nos impulsiona para construir modelos que consideram uma componente estocástica adicional. Assim, nossa nova equação modelo pode ser compreendida na forma

$$x(t) = \phi_1 x(t-d) + \phi_2 x(t-2d) + N(\mu, \sigma),$$

onde ϕ_1 e ϕ_2 são coeficientes relacionados às medidas respectivas com atrasos $t-d$ e $t-2d$, e $N(\mu, \sigma)$ representa valores aleatório gerados seguindo uma distribuição normal (ou gaussiana) com média μ e desvio padrão σ.

Como os parâmetros d e m alteram a acuidade da predição realizada sobre uma dada série temporal, em próximo passo precisamos entender os modelos *Auto-Correlation Function* (**ACF**, ou Função de Auto-Correlação), *Average Mutual Information* (**AMI**, ou Informação Mútua Média) e *False Nearest-Neighbors* (**FNN**, ou Falsos Vizinhos Próximos).

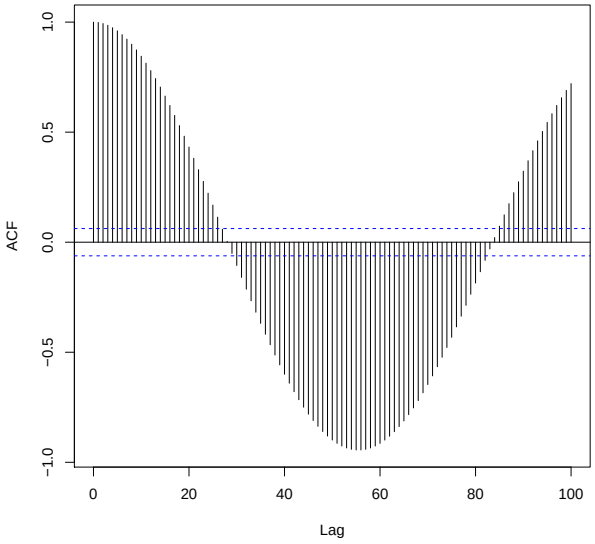


Figura 7: **ACF** medida para a função seno considerando $d \in [0, 100]$. O tracejado em azul indica o intervalo de confiança para que a função seja considerada temporalmente independente.

A **ACF** nos indica a medida de covariância sobre a mesma matriz X dada por $X \cdot X^T$ (produto escalar, com X^T transposta de X) e neste contexto é interpretada como uma medida de dependência temporal entre uma medida $x(t)$ com seu valor nos demais passos temporais. Deste modo, quanto mais próximo ao valor de ± 1, maior a dependência temporal. Por outro lado, quanto mais próximo a zero (medido de acordo com o intervalo de confiança) menor

a dependência temporal. Vale lembrar que o cenário do **TET** é aplicável no caso de existir dependência temporal.

Ao considerarmos a função seno, na figura 7 temos o valor de **ACF** (eixo vertical) calculado para cada um dos atrasos temporais (*lag*, no eixo horizontal). Notamos alta dependência temporal mesmo para 20 deslocamentos temporais, cujos valores de **ACF** são maiores que o intervalo de confiança (linha tracejada em azul), indicando que pode existir uma dependência temporal e, portanto, o modelo de **TET** pode nos prover bons resultados.

Em segundo momento avaliaremos a métrica **AMI**. Esta medida se propõe a estabelecer um critério tal que seja possível determinar um valor ótimo de d que maximize a informação recuperada através da dependência temporal. Em princípio poderíamos escolher um valor de atraso d que esteja associado a uma auto-correlação de valor zero. Neste atraso, a correlação é entendida como sendo nula e um modelo linear poderia ser traçado entre $x(t)$ e $x(t-d)$, como apontado na literatura [5]. No entanto, a **ACF** assume relação linear entre duas variáveis, o que restringe sua utilização (devido à sua relação ser dada pelo produto escalar entre os valores da matriz M que contém os valores de $x(t)$ presente na série temporal medida).

Como a informação de um sistema pode ser medida de acordo com a entropia do estado em que a medida foi feita na forma

$$H(S) = -\sum_i P_s(s_i)\log[P_s(s_i)], \tag{10}$$

onde $H(S)$ nos retorna a entropia de Shannon [28], S representa a configuração atual do sistema dado que nela temos os estados $s_i \in S$ cuja probabilidade de ocorrência individual é $P_s(s_i)$.

Com esta medida, podemos definir $[s,q] \equiv [x(t), x(t+d]$ tal que seja possível analisar a incerteza (entropia) da medida q (medida feita sobre o atraso temporal d) dado que tenhamos medido s no instante atual t através da relação condicional [5]

$$\begin{aligned} H(Q|s_i) &= -\sum_j P_{(q|s)}(q_j|s_i)\log[P_{(q|s)}(q_j|s_i)] \\ &= -\sum_j \left[\frac{P_{sq}(s_i,q_j)}{P_s(s_i)}\right]\log\left[\frac{P_{sq}(s_i,q_j)}{P_s(s_i)}\right], \end{aligned} \tag{11}$$

onde $P_{(q|s)}(q_j|s_i)$ é a probabilidade de que a medida q seja q_j dado que a medida de s é s_i, e $P_{sq}(s_i,q_j)$, a probabilidade conjunta de ocorrência dos eventos s_i e q_j. A última igualdade decorre da relação bayesiana de probabilidade condicional $P(A|B) = P(A\cap B)/P(B)$.

Como a medida é feita sobre cada um dos estados conhecidos s_i, a incerteza média de que $x(t)$ ocorra para $x(t+d)$ é dada pela média de $H(Q|s_i)$ sobre todo valor s_i de modo que

$$H(Q|S) = \sum_i P_s(s_i)H(Q|s_i) = H(S,Q) - H(S), \tag{12}$$

tal que

$$H(S,Q) = -\sum_{i,j} P_{s,q}(q_j,s_i)\log[P_{q,s}(q_j|s_i)], \tag{13}$$

representa a entropia de S condicionada à ocorrência de Q e $H(S)$ a entropia marginal do estado S.

Por fim, definimos **MI** (*Mutual Information*) como

$$\begin{aligned} I(Q,S) &= H(Q) - H(Q|S) \\ &= H(Q) + H(S) - H(S,Q) = I(S,Q). \end{aligned} \tag{14}$$

Podemos agora interpretar o que esta medida representa em termos da entropia dos estados do sistema. Pela equação 14 notamos que $I(Q,S)$ providencia uma medida de acurácia de quanta informação uma medida sobre a variável x (na forma da entropia $H(Q)$) é informativa para uma segunda medida de x (na forma $H(S)$), onde valores reduzidos de informação mútua representam bons candidatos para o parâmetro d, já que pouca informação existe de diferente entre o estado $x(t)$ e $x(t+d)$ (resultado de $H(S,Q)$ numericamente grande relativo aos valores

de $H(S)$ e $H(Q)$). De modo menos rigoroso podemos afirmar que o primeiro mínimo local agrega o melhor atraso d dado que a difusão de entropia para sistemas caóticos é sempre positiva [5].

Utilizando uma série temporal senoidal, obtemos a **AMI** (figura 8) conforme variamos o atraso temporal *lag* (d, representado pelo eixo horizontal) apresentada na figura, cujo primeiro mínimo local é dado pelo lag (atraso) temporal $d = 4$. Ou seja, um atraso de $d = 4$ passos temporais é ideal para a representação deste sistema.

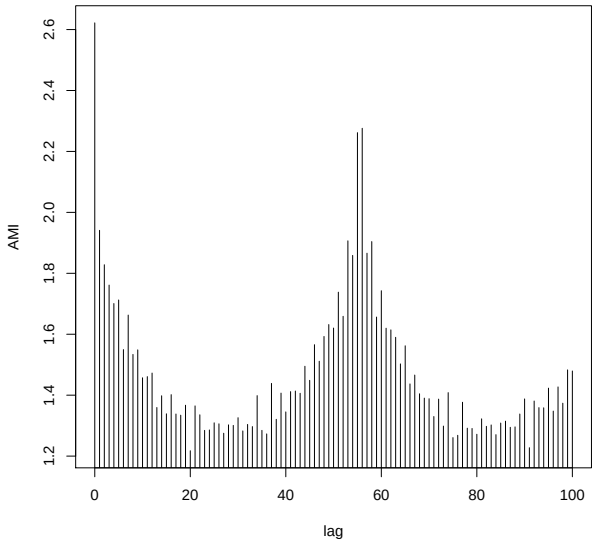


Figura 8: **AMI** medida para uma função seno com atrasos temporais $d \in [0, 100]$.

Por fim descreveremos a medida **FNN**. Nela, temos um critério para escolher valores de m tal que este número de dimensões seja ótimo para desdobrar o espaço unidimensional (em relação a t) em uma representação multivariada de diversas medidas de x em distintos instantes de tempo [12]. Aplicado à evolução temporal da senoide, podemos re-interpretar como o número de atrasos temporais, com $d = 4$, que precisamos considerar para que a reconstrução da série seja feita.

O método apresentado como **FNN** considera que atratores são compactos no espaço de fase e que, portanto, "ganham vizinhos"neste espaço conforme o avanço temporal. Neste cenário, a distância euclidiana é interessante de ser avaliada entre dois determinados pontos. Tomando que o desdobramento do espaço $x(t)$ foi feito com um atraso temporal d para m dimensões, a distância para o $r-$ésimo vizinho mais próximos de um ponto de consulta fixo $x(n)$ é dada por

$$R_m^2(n,r) = \sum_{k=0}^{m-1} [x(n+kd) - x^{(r)}(n+kd)]^2, \qquad (15)$$

Caso consideremos uma nova dimensão $m+1$, podemos manter fixo o vizinho $x^{(r)}$ e determinar a nova distância euclidiana

$$R_{m+1}^2(n,r) = \\ = R_d^2(n,r) + [x(n+md) - x^{(r)}(n+md)]^2, \qquad (16)$$

uma vez que a inclusão de uma nova dimensão apenas considera uma nova componente $[x(n+md) - x^{(r)}(n+md)]^2$ inclusa pela última dimensão m adicionada.

Com esta formulação, um vizinho falso é designado quando as relações

$$\left[\frac{R_{m+1}^2(n,r) - R_m^2(n,r)}{R_m^2(n,r)}\right]^{1/2} > R_{tol}; \qquad (17)$$

$$\frac{|x(n+md) - x^{(r)}(n+md)|}{R_d(n,r)} > R_{tol}. \qquad (18)$$

são válidas dado um limiar R_{tol}.

Neste sentido, podemos considerar apenas uma dimensão e, com seu incremento, contabilizar quantos vizinhos considerados são falsos. Como originalmente reportado [12], para sistemas atratores com alta dimensionalidade (geradores de números aleatórios, e.g.) o incremento de pontos resulta na divergência do número N de falsos vizinhos próximos, uma vez que mesmo altas dimensões não são capazes de agregar todos os possíveis falsos vizinhos (que podem ter sido projetados sobre o ponto

de consulta), fazendo com que seu número possa ser uma função não limitada conforme aumentamos m.

Ao considerarmos a série senoidal, obtemos o percentual de falsos vizinhos apresentado na figura 9. Nela, podemos concluir que $m = 4$ representa uma dimensão de inserção suficientemente boa ($\%$ **FNN** $\approx 8e(-3)$) tal que o percentual de falsos vizinhos próximos é próximo a zero.

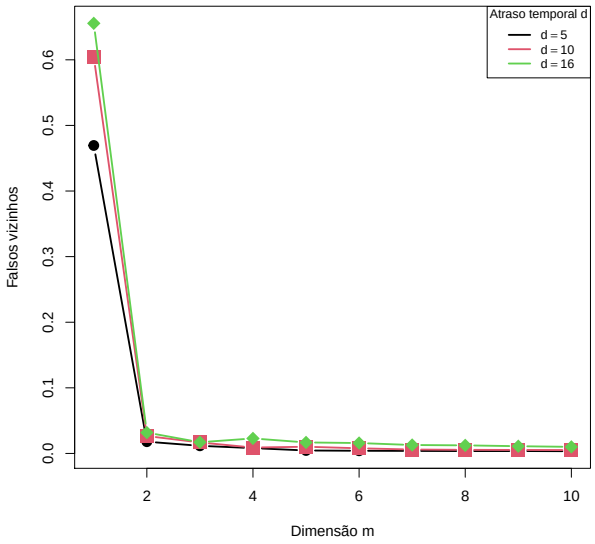
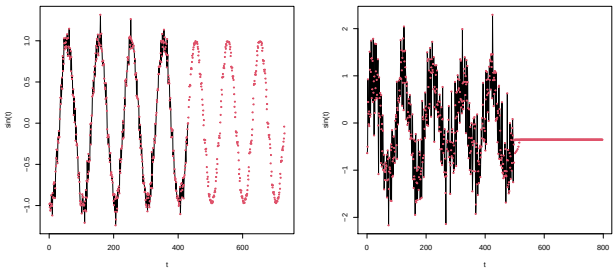


Figura 9: **FNN** calculado para função seno aplicado sobre as dimensões incrementais de m até $m = 10$. Consideramos diferentes atrasos temporais $d = 5, 10, 16$ em círculos pretos, quadrados vermelhos e losangos verdes, respectivamente.

Com estes resultados, podemos determinar os parâmetros m, d que melhor representam uma dada série temporal. Assim, determinamos um d ótimo que reduz a perda de informação, o qual será fixo para determinarmos a menor dimensão m cujo espaço de fase é melhor desdobrado, reduzindo a superposição de estados. Finalmente podemos utilizar de um algoritmo de regressão (caso do **DWNN**) para realizar uma análise preditiva da série.

Como exemplo, podemos gerar uma evolução temporal em senoide em conjunto a um ruído com distribuição normal com média zero. Utilizando a metodologia descrita no último parágrafo, notamos um melhor poder preditivo na figura 10 (onde os pontos vermelhos são predições) quando comparado com a análise unicamente utilizando **DWNN** na figura 6.



(a) (b)

Figura 10: Previsão senoide com ruído com uso de **AMI**, **FNN**, **DWNN** com $\sigma = 0.1$ (a) e $\sigma = 0.5$ (b). Com estes algoritmos, conseguimos prever (em pontos vermelhos) o comportamento ruidoso para senoide com pouco ruido, mas a predição é comprometida quando sua aleatoriedade é alta. Nelas, usamos o valores $d = 3$ e $m = 2$.

Na figura 10 temos duas curvas senoidais em função do tempo com adição de ruído gaussiano com desvio padrão $\sigma = 0.1$ (a) e $\sigma = 0.5$ (b), e notamos que o modelo utilizado prevê o comportamento bastante acurado para séries temporais ruidosas, no entanto não apresenta muita acuidade quando na presença de ruídos muito intensos.

Desta forma, passamos para uma nova abordagem de modelo preditivo para séries temporais chamada *Empirical Mode Decomposition* (**EMD**, ou Decomposição por Modo Empírico), a qual busca resolver o problema emergente da análise acima mencionada.

Por construção, nossa série temporal é dada por uma componente senoidal com o avanço do tempo somada a um ruído aleatório com distribuição gaussiana em torno da média zero. Assim, é razoável imaginar que a mesma possa ser reconstruída através da decomposição de duas

funções temporais $F(\tau)$ e $G(\tau)$ tais que $x(t) = F(\tau) + G(\tau)$, onde $F(\tau)$ é chamada componente determinística e $G(\tau)$, componente estocástica. A variável τ indica dependência temporal com possíveis atrasos.

A **EMD** considera a decomposição de uma dada série temporal em distintas *Intrinsic Mode Functions* (**IMF**, ou Funções de Modo Intrínseco) e apresenta-se como um possível modelo de separação entre **IMF**s predominantemente determinísticas e **IMF**s predominantemente estocásticas [25].

Em primeiro lugar devemos entender como cada **IMF** é construída. Dada uma série temporal, uma **IMF** é definida como uma função que possa ser descrita como um processo gaussiano estacionário e que todo ponto deve apresentar média zero dentro do envelope limitado pelo máximo e mínimo neste ponto [9].

A definição acima pode ser melhor compreendida na figura 11, onde aplicamos ruídos intensos em distribuição gaussiana $N(\mu = 0, \sigma = 0.5)$. Os pontos em vermelho representam a média móvel entre o máximo e mínimo da função que envelopa o toda a série e é dada por

$$y(t) = \alpha x(t) + \beta y(t-1), \tag{19}$$

onde $y(t)$ representa o valor médio da série no instante t e α, β parâmetros variáveis. Em nosso caso particular para a construção da figura 11 utilizamos $\alpha = 0.9$ e $\beta = 0.1$.

A função recursiva acima pode ser entendida como uma média móvel ponderada por série de potências de $(1-\alpha)$ (apesar de sua forma linear) devido às relações de recursão

$$\begin{aligned} y_0 &= x_0 \\ y_1 &= \alpha x_1 + (1-\alpha)x_0 \\ y_2 &= \alpha x_2 + [1-\alpha][\alpha x_1 + (1-\alpha)x_0] \\ &= \alpha x_2 + \alpha(1-\alpha)x_1 + (1-\alpha)^2 x_0 \\ &\dots \end{aligned}$$

onde impomos que $\alpha, \beta \in [0,1]$ e que $\beta = 1 - \alpha$ (complementar).

A fórmula geral para um instante de tempo t com k médias calculadas anteriores é dado por

$$\begin{aligned} y_t = \alpha\big[x_t + (1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-2} + \dots \\ + (1-\alpha)^k x_{t-k}\big] + (1-\alpha)^{k+1} y_{t-(k+1)}, \end{aligned} \tag{20}$$

que nos permite entender que o peso associado à média y_{t-i} é dado por $\alpha(1-\alpha)^i$, evidenciando a dependência por potências.

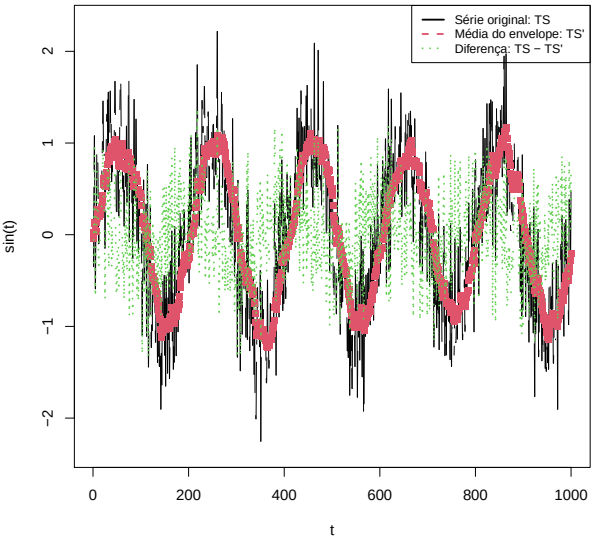


Figura 11: Função seno com ruído gaussiano com $\mu = 0$ e $\sigma = 0.5$ original (TS, traço contínuo preto) com média móvel ponto a ponto TS' (tracejado vermelho) e diferença entre a série original $TS - TS'$ e a média móvel (pontos verdes).

Ao tomarmos a diferença entre o valor real da série $x(t)$ e a média do envelope neste ponto $y(t)$ (representado pelo tracejado verde na figura 11, onde TS indica a série original e TS' a média móvel), podemos analisar a propriedade *zero-crossing* mencionada anteriormente. Quanto maior o número de cruzamentos desta função através da linha de valor nulo, maior a tendência senoidal presente

na série original. Já um processo gaussiano estacionário implica um retorno a um ponto médio no espaço pela série temporal original. A descrição da estacionariedade de um processo será explicada com mais detalhes futuramente.

Com estas duas condições, as **IMF**s descrevem modos de oscilação dominantes nos dados originais. Impondo que cada senoide que possivelmente componha os dados originais apresentem frequências de oscilação distintas, para cada intervalo temporal teremos oscilações intrínsecas a cada escala de tempo característica (duas curvas seno com frequências distintas não completam suas fases em um mesmo deslocamento de tempo, e.g.).

A determinação de cada **IMF** decorre da construção de uma nova série dada pela diferença entre a média móvel e os dados da série anterior. Desta forma, sendo m_1 a primeira curva de média móvel encontrada (exemplo da curva em vermelho na figura 11),

$$h_1 = X(t) - m_1 \tag{21}$$

nos fornece h_1, chamada de primeira componente ou IMF_1.

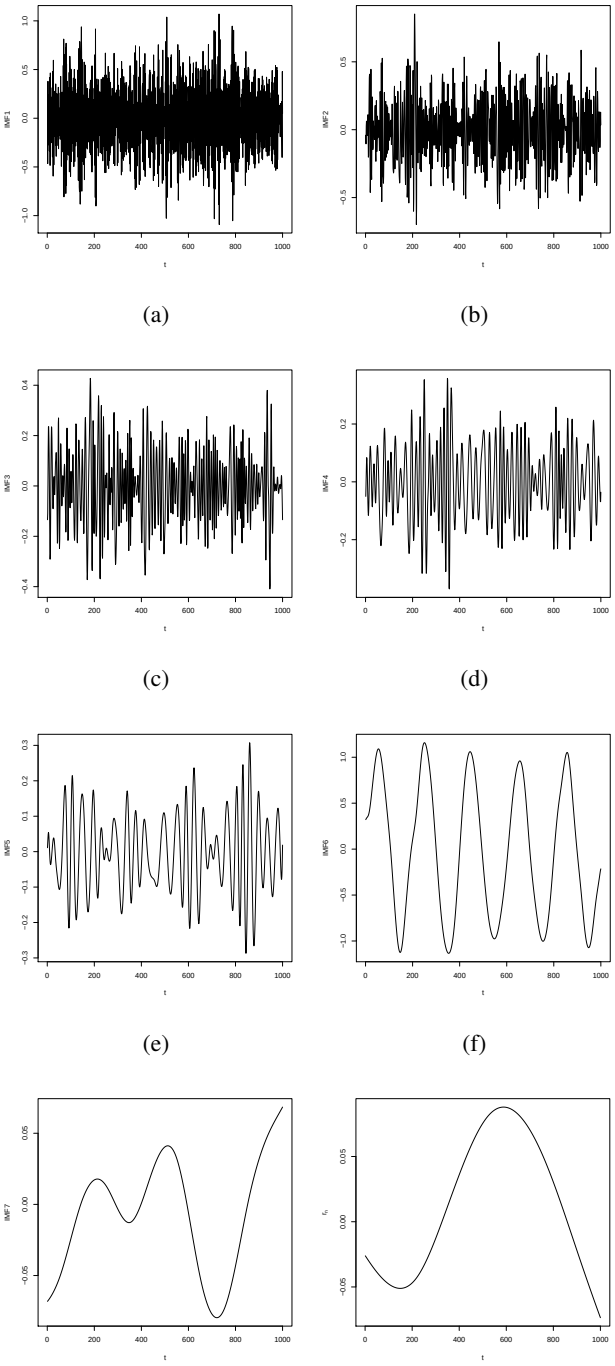
De modo análogo, podemos tratar h_1 como o novo dado original e, sendo a nova média móvel m_{11}, temos

$$h_{11} = h_1 - m_{11} \tag{22}$$

representando a segunda componente intrínseca, ou $\textbf{IMF}_2$.

O processo é terminado quando não existem mais oscilações capazes de completar um ciclo em sua fase. A esta última componente damos o nome resíduo r_n. Neste sentido, sendo que existe um acúmulo de resíduos a cada **IMF** encontrada, de modo geral temos que uma série original $X(t)$ pode ser generalizada como

$$X(t) = \sum_{i=1}^{n} c_i + r_n \tag{23}$$



(a) (b) (c) (d) (e) (f) (g) (h)

Figura 12: Decomposição em **IMF**s para a curva senoidal com ruído gaussiano $\sigma = 0.5$. Conforme a ordem da diferença é incrementada, notamos a obtenção de comportamentos mais determinísticos. Em (h) representamos o resíduo da decomposição.

onde c_i representa a $i-$ésima componente intrínseca já incluso acúmulos de seu resíduo r_1 (tal que $c_1 = r_1 + X(t)$) e r_n o resíduo que acompanha a última decomposição [9].

Os resultados apresentados na figura 12 nos motiva a perceber a existência de componentes mais ruidosas que outras. Por exemplo, quando comparamos a primeira $\mathbf{IMF}_1$ em (a) com a $\mathbf{IMF}_6$ em (f), notamos que esta última se expressa com dependência senoidal suave com o tempo. Com isso, tal decomposição potencialmente carrega informação útil sobre o número de componentes necessárias para que sejam desagregados os comportamentos estocástico e determinístico. De fato, estas componentes podem ser expressas no espaço de Fourier e são bons delimitantes das componentes $F(\tau)$ e $G(\tau)$.

Quando aplicamos a transformada de Fourier (**FT**) sobre uma dada série, ela é responsável por decompor as frequências dos modos normais das curvas senoidais que representam a série original por meio da função

$$\mathcal{F}(X(k)) = \sum_{n=1}^{N} a_n e^{-i(2\pi k)n/N}, \tag{24}$$

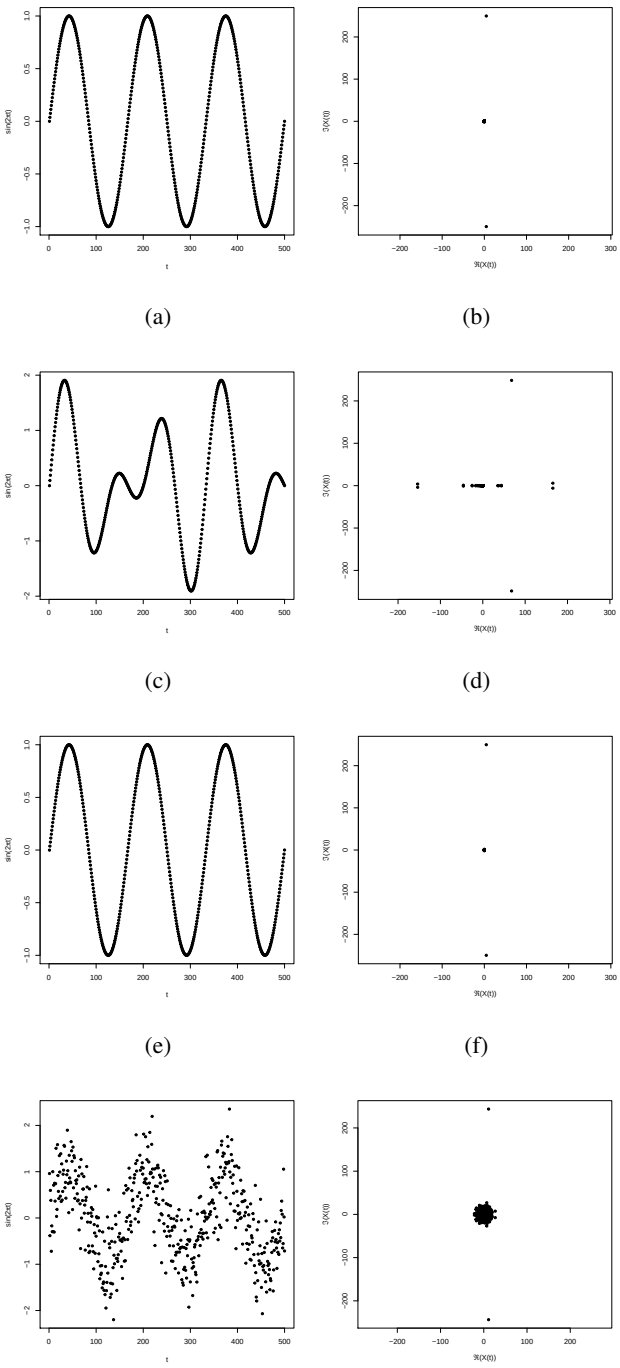
tal que $\omega_k = 2\pi k$ representa a frequência característica da curva com número de onda k em uma série de tamanho N. O espaço de frequências k é denominado **espaço de Fourier**.

Quando unimos este conceito à decomposição em **IMF**s [25], podemos passar da representação espaço-tempo (**IMF**, com $h_i(t)$) para o espaço das frequências a elas associadas (**FT**, com $\mathcal{F}(h_i(t))$) de modo que $C_j(t) = c_{j,1}, c_{j,2}, \ldots, c_{j,\tau}$ representa agora um conjunto de coeficientes complexos $c_{j,k}$ tais que

$$c_{j,k} = \sum_{t=1}^{T} h_j(t) e^{-i(2\pi k)t/T}, \tag{25}$$

com $k \in 1, 2, \ldots, T$.

Como as frequências no espaço de Fourier seguem coor-



(a) (b) (c) (d) (e) (f) (g) (h)

Figura 13: Decomposição no espaço de Fourier para curvas em forma senoidal dadas por $\sin(2\pi t)$ (a) e (b), $\sin(2\pi t) + \sin(3\pi t)$ (c) e (d), $\sin(2\pi t) + N_1$ (e) e (f), $\sin(2\pi t) + N_2$ (g) e (h), onde N_1, N_2 representam ruídos gaussianos, respectivamente com $\sigma = 5e-4$ e $\sigma = 0.5$.

denadas distribuídas nos eixos imaginário e real dado que

$$\rho e^{-i\theta} = \rho[\cos\theta - i\sin\theta] \tag{26}$$

de acordo com a fórmula de Euler, então para uma dada série podemos aplicar a equação 24 e obter as componentes $\mathrm{Im}\{\theta\}$ e $\mathrm{Re}\{\theta\}$. Para algumas combinações de curvas senoidais, apresentamos na figura 13 suas representações no espaço-tempo e no espaço de Fourier.

Para as primeiras duas senoides geradas (13 (a) e (c)), utilizamos respectivamente uma função $\sin(2\pi t)$ e a combinação $\sin(2\pi t) + \sin(3\pi t)$. Notamos que a amplitude (tamanho dos eixos) não altera no espaço de frequências (13 (b) e (d)), porém os pontos são rotacionados pelo ângulo proporcional à frequência de cada componente somada.

Mais precisamente, o ângulo é dado pela equação

$$\theta = \tan^{-1}\left[\frac{\sin\theta}{\cos\theta}\right], \tag{27}$$

que pode ser generalizado na forma [25]

$$\theta(h_j(t)) = \tan^{-1}\left[\frac{\mathrm{Im}\{C_j(t)\}}{\mathrm{Re}\{C_j(t)\}}\right]. \tag{28}$$

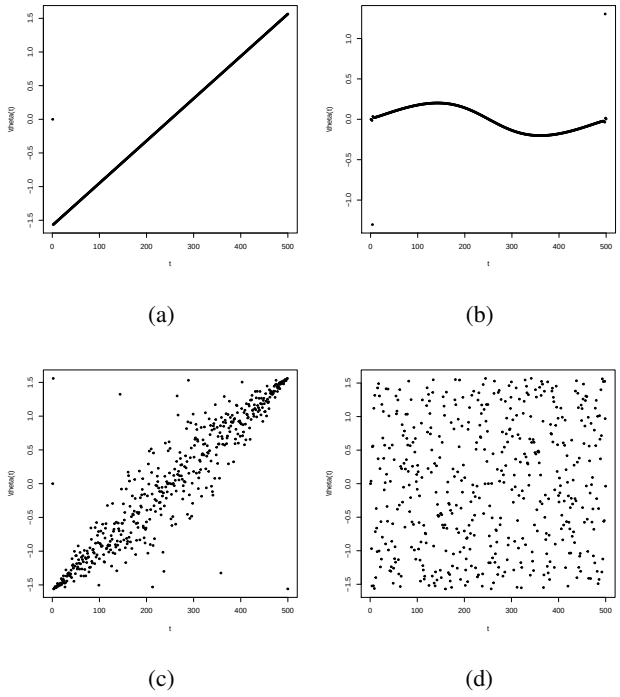
Por outro lado, quando consideramos a função $\sin(2\pi t)$ original em conjunto com um ruído, obtemos as figuras 13 (e) e (g). Nelas, em ordem, adicionamos um ruído aleatório gerado pela distribuição gaussiana $N(\mu = 0, \sigma = 5(10^{-4}))$ e $N(\mu = 0, \sigma = 0.5)$. Na primeira, o ruído apresenta baixa dispersão em torno da média, o que mantém a curva temporalmente bem comportada e com representação no espaço de frequências (13 (f)) similar à curva seno original (13 (b)). Já para a adição de ruído mais disperso, a função passa a apresentar caráter mais estocástico (claramente visto em 13 (g)). Seu mapeamento no espaço de Fourier (13 (h)) aproxima-se da original (13 (b)), mas os pontos não apresentam núcleo bem definido.

Neste último caso, fica mais claro o que chamaremos de **congruência de fase**. No espaço de frequências, o incremento temporal rotaciona em um ângulo θ (obtido pela equação 27) um vetor de norma ρ (representado pela equação 26). Com isso, pontos ao centro dizem respeito à combinação de curvas com alta frequência (cuja amplitude se soma de forma destrutiva) e pontos distantes do centro referem-se a instantes de tempo cuja combinação resulta na construtividade entre as séries consideradas, uma vez que a amplitude ρ sofre interferência construtiva.

Assim, quando vários pontos convergem para uma região pequena no espaço de Fourier, temos a congruência de fase, pois encontramos um instante a partir do qual, após um período de 2π, novas medidas convergem para o mesmo ponto. De forma mais visual, a figura 13 (b) representa uma congruência perfeita de fases (toda medida é centrada na origem) e a figura 13 (h) representa uma congruência mais precária, cujas medidas oscilam em torno do centro, mas não convergem para ele.

Uma nova análise que pode ser feita se dá através da representação desta congruência por meio da variação de ângulos de deslocamento associado a cada ponto no espaço de Fourier.[25] Utilizando da equação 28 apresentamos na figura 14 as mesmas 4 curvas senoidais utilizadas na figura 13, porém representadas de acordo com suas fases em função do tempo.

Por construção sabemos que as figuras superiores (14 (a) e (b)) representam curvas determinísticas e as inferiores, (14 (c) e (d)), uma combinação com componente estocástica por conta da adição de ruído. Fato interessante ocorre para a sub-figura (c), a qual mesmo contendo uma componente gaussiana aleatória com pouco desvio (i.e. $\sigma \approx 10^{-4}$), o mapeamento para a mudança de fase é expressivo na indicação da presença de elementos aleatórios. Apesar disso, também é possível notar a convergência para uma reta análoga à sub-figura (a). Nela fica evidente a congruência de fase, que pode ser compreendida como a variação suave de ângulo entre uma medida no instante



(a) (b)

(c) (d)

Figura 14: Congruências de fase para as 4 senoides construídas para a figura 13. São elas: $\sin(2\pi t)$ (a), $\sin(2\pi t)+\sin(3\pi t)$ (b), $\sin(2\pi t)+N_1$ (c), $\sin(2\pi t)+N_2$ (d) onde N_1, N_2 representam ruídos gaussianos, respectivamente com $\sigma = 5e-4$ e $\sigma = 0.5$. Notamos que quando adicionado um ruído, mesmo com baixa dispersão, a congruência de fase (c) é drasticamente espalhada.

t em relação a medidas subsequentes em $t-\delta$ e $t+\delta$ (com δ representando um passo temporal pequeno).

Com isso, conseguimos determinar quantas e quais são as componentes predominantemente determinísticas e aquelas com maior tendência estocástica. Isso nos permite construir nossa análise preditiva para a série original através do modelo *Auto-Recursive Integrated Moving-Average* (**ARIMA**, ou Modelo Auto-Recursivo Integrado com Média Móvel).

O modelo **ARIMA** [7] considera que uma dada série temporal é composta de componentes determinísticos, que podem ser modelados por auto-regressão (ou outros modelos de sistemas dinâmicos, que consideram dependências de eventos passados suficientes para sua predição de evoluções temporais futuras) somados a um número de componentes estocásticos, modelados por meio de ferramentas de aprendizado estatístico (como o caso de cadeias de markov e média móvel). Associados a eles temos possivelmente um número não nulo de componentes integradas, as quais apresentam um nível de dependência temporal entre duas dadas variáveis.

Em equação, podemos denotar este modelo (em particular o modelo **ARMA**, o qual desconsidera termos integrados) na forma[25]

$$X(t) = a_1X(t-1) + \ldots + a_pX(t-p) + \\ + e(t) + b_1e(t-1) + \ldots + b_qe(t-q), \qquad (29)$$

onde $a_1, a_2, \ldots, a_p$ são coeficientes associados à decomposição determinística considerando p atrasos temporais, e $b_1, b_2, \ldots, b_q$, coeficientes da decomposição estocástica, levando em conta q atrasos temporais com distribuições de ruído dadas pela função $e(.)$.

Neste momento temos o número de componentes estocásticos e determinísticos dados pelas **IMF**s via congruência de fase, o atraso temporal d e o número de dimensões inseridas m que melhor descrevem a série temporal original e o modelo **DWNN** de regressão para aprender a evolução no tempo das componentes determinísticas. Como métodos finais, apresentaremos os modelos de aprendizado para componentes estocásticas *Partial Auto-Correlation Function* (**PACF**, ou Função de Auto-Correlação Parcial) em conjunto com a análise de estacionariedade (importante para utilização de modelos estocásticos, como previamente apontado neste estudo [9]).

A hipótese de estacionariedade das componentes estocásticas extraídas da série original reflete na estabilidade dos momentos estatísticos ao longo do tempo.

Atrelados a este efeito sobre os momentos, séries não estacionárias podem apresentar correlação de longo alcance (i.e., séries cuja dependência do atraso temporal seja grande) e períodos transientes, cuja previsibilidade pode ser prejudicada [30]. Os momentos de uma distribuição são definidos como [29]

$$\mu_n = \langle x^n \rangle = \int_{-\infty}^{\infty} x^n \rho(x) dx \tag{30}$$

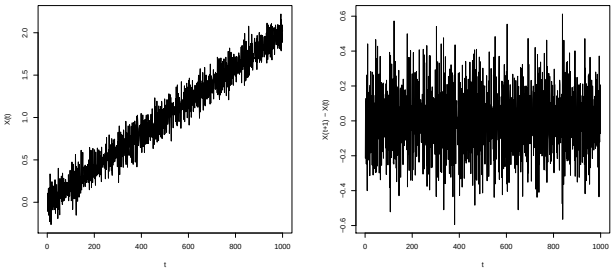
tal que $\rho(x)$ representa a distribuição sobre a qual desejamos encontrar o $n-$ésimo momento.

Cada momento expressa uma determinada métrica relevante para uma distribuição $\rho(x)$. Por exemplo, os primeiros 3 momentos μ_1, μ_2, μ_3 representam, respectivamente, a média, a variância e a assimetria (conhecida como *skewness*).

Podemos compreender se uma série é estacionária de ordem c caso a diferença entre o $c-$ésimo momento de intervalos da série original são próximos a zero. Quando temos o caso particular para os dois primeiros momentos, denominamos estacionariedade *fraca*. Para estacionariedade dita *forte* é necessário que a probabilidade conjunta de secções distintas da série original seja a mesma, indicando independência temporal do espaço de inserção (*embedding space*) estatísticas [30]. Neste estudo trataremos apenas de séries estacionárias com ordem 1.

Para séries com média e/ou variância com tendência não estacionária, podemos calcular a diferença ponto a ponto do valor atual com o anterior. Esta diferença é suficiente para cancelar a tendência crescente (ou decrescente) de uma série, uma vez que preserva a variação entre dois pontos adjacentes e translada a média para que pertença ao intervalo entre $x(t+1)$ e $x(t)$. Este efeito pode ser visualizado na figura 15.

Quando impomos estacionariedade sobre a série original (via cálculo da diferença entre eventos subsequentes, por exemplo), podemos determinar a dependência temporal



(a) (b)

Figura 15: Serie não estacionária com desvio de ordem 1 (a) e construção estacionária com diferença ponto a ponto (b).

que eventos com atraso no tempo exercem sobre o evento atual de modo mais preciso. Para isto, utilizaremos a função **PACF**.

A **PACF** apresenta construção análoga à **ACF**, uma vez que quantifica a correlação (análise de covariância) entre duas medidas de uma série temporal com atraso. No entanto a **ACF** é responsável por analisar a correlação para o atraso d existente entre $x(t)$ e todas as demais medidas em $x(t-d)$, já a **PACF** cujo atraso é d quantifica a correlação entre $x(t)$, $x(t-d)$ e todas os passos intermediários, ou sejam entre $x(t), x(t-1), \ldots, x(t-d)$ [14].

Assim, sendo $(x_t | x_{t-1}, x_{t-2}, \ldots, T_{t-d+1})$ a medida de x_t dado que medimos $x_{t-1}, x_{t-2}, \ldots, x_{t-d+1}$ definido como $x_{(t-1,\ldots,t-d+1)}$ e $(x_{t-d} | x_{t-1}, x_{t-2}, \ldots, T_{t-d+1})$ a medida de x_{t-d} dado que medimos $x_{(t-1,\ldots,t-d+1)}$, temos que

$$PACF(x_t, d) = \left[x_t | x_{(t-1,\ldots,t-d+1)} \right] \cdot \left[x_{t-d} | x_{(t-1,\ldots,t-d+1)} \right]. \tag{31}$$

Neste sentido, a **ACF** nos retorna o atraso temporal d referente à componente determinística e **PACF**, o atraso d referente à estocástica, uma vez que considera a ocorrência condicional de uma medida $x(t)$ após d passos da medida $x(t-d)$.

Com isso concluímos a metodologia que será aplicada sobre a análise da série temporal escolhida, cujos resultados serão discutidos na seção V-B.

V. Resultados

A. IBL

A partir da motivação inicial mencionada na introdução a respeito dos estilos de governo, para o início deste trabalho precisamos estabelecer alguns dos parâmetros relevantes ao longo deste estudo inicial.

Se considerarmos os regimes políticos em dois casos extremos em autocracia e democracia, podemos avaliar a evolução histórica de ambos. Os dados coletados consideram a estrutura dos regimes de 1900 a 2017.

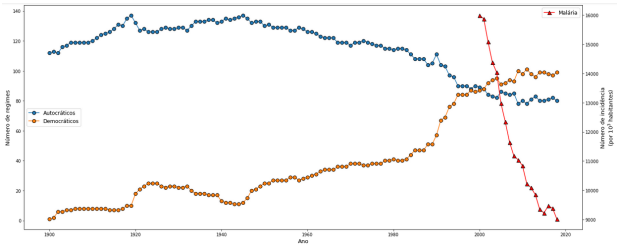


Figura 16: Evolução anual do número de democracias e autocracias (escala à esquerda) pelo mundo. No mesmo gráfico, a escala à direita considera o número de casos de malária em escala global por 1000 habitantes.

É interessante perceber que até 2000 o número de autocracias supera o número de democracias ao redor do mundo. Neste ano, no entanto, o cenário se reverte e o número de democracias passa a ser maior que o número de autocracias. Neste sentido, podemos especular que devem existir dados que correlacionam fortemente com o estilo de governo de um determinado país. Quando consideramos um país em específico, em particular o Brasil, é possível analisar o desenvolvimento do processo democrático através de métricas do índice político.

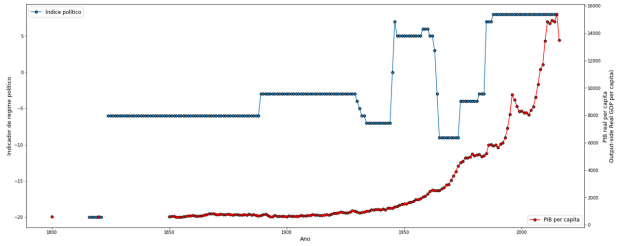


Figura 17: Evolução anual do índice político brasileiro entre os anos de 1900 a 2017 (escala à esquerda). Em conjunto, temos o crescimento anual do PIB (Produto interno Bruto) nacional na escala à direita.

Esse índice considera, entre outras nuances, [13] a liberdade de expressão de uma população e a forma de obter o poder político, onde -10 indica um governo autocrático hereditário, onde a governança é transferida se o consentimento da população, e +10, governos democráticos liberais, onde existe governantes escolhidos através de eleições diretas e representativas.

Assim, notamos que existe uma correlação entre o crescimento do Produto Interno Bruto (PIB) brasileiro e o índice político, em especial após o fim da ditadura militar (1985) e um maior crescimento do PIB.

Motivados por essa correlação, podemos tomar métricas econômicas e relacionar predições do índice político.

As métricas econômicas foram retiradas da base de dados PWT91[3]. Nesta tabela, encontramos índices de entrada e saída financeira, além de índices de produtividade de 1950 a 2017. Como restrição do escopo deste trabalho, apenas as métricas referentes ao Brasil serão consideradas.

Ao total, o conjunto de dados utilizados apresenta 52 colunas representando PIB nacional, PIB *per capita*, taxas de produtividade e empregabilidade, entre outras. Para que estes dados sejam mais tratáveis, utilizaremos a análise de **PCA**. Neste caso em específico seria como responder

[3]Dados em:
https://www.rug.nl/ggdc/productivity/pwt/?lang=en

a pergunta: quantos e quais são as variáveis, dentre as 52, que apresentam maior variância e, com isso, melhor segregam a base de dados?

Ao aplicar esta análise sobre nossa base de dados PWT91, cada uma destas componentes têm associado um autovalor representativo da variância dos dados sobre ela. Podemos então mostrar em gráfico a função cumulativa de autovalores em um gráfico de *scores*. Foram consideradas apenas 34 das 52 colunas, uma vez que parte delas representam variáveis de identificação ou lógicas (0 ou 1).

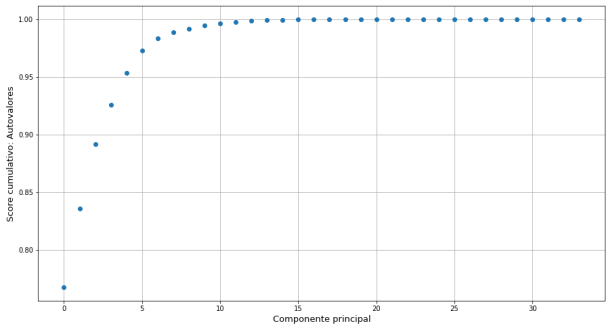


Figura 18: Função cumulativa dos autovalores associados a cada um dos autovetores da matriz de covariância dos dados da base PWT91. Restringimos o gráfico entre 34 das colunas da base original.

Como cada autovalor representa a variância dos dados, ao restringir todo o conjunto de dados às três primeiras componentes, conseguimos próximo a 90% de representatividade. Neste sentido, temos mais claro o motivo do nome PCA, uma vez que conseguimos reduzir toda nossa base em apenas 3 componentes principais que já são responsáveis por reter aproximadamente 90% da informação original.

Já com o uso dos autovetores, podemos rotacionar cada um dos dados para que os eixos vertical e horizontal correspondam a cada uma das componentes principais. Por fim, como cada métrica está associada a um ano e a cada ano temos um índice político associado (em escala discreta de -10 a +10, como mencionado anteriormente), podemos representar nossas três componentes nos gráficos que seguem, onde cada cor indica o índice político.

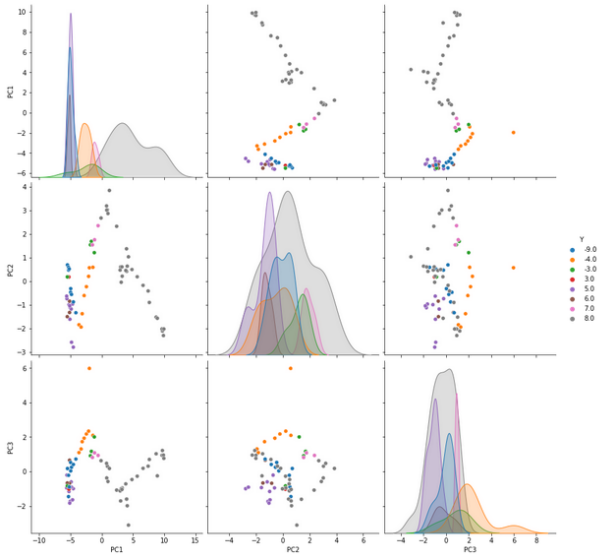


Figura 19: Representação das três componentes principais do banco de dados PWT91. Na diagonal, apresentamos os histogramas da frequência relativa de cada indicador mostrado através das cores indicadas na legenda. Por exemplo a cor azul representa indicadores fortemente autocráticos (-9) e a cor cinza, fortemente democrático (+8).

Observando a representação em cores distintas para cada indicador político, é razoável estabelecer que existam hiperplanos cujo conteúdo seja capaz de prescrever um indicador a um determinado ano com certa acurácia. No entanto

Outro resultado interessante possível de se obter através do algoritmo de classificação **KNN** refere-se à predição dos índices respectivos aos anos de 2016 e 2017, uma vez que os dados econômicos foram coletados neste período, o que não ocorreu com os indicadores políticos. Neste sentido, utilizando de três componentes principais e com $K = 7$ vizinhos mais próximos (capaz de minimizar o erro absoluto), para 2016 encontramos o valor $(+5 \pm 3)$ e para 2017, o valor $(+6 \pm 3)$, indicando uma possível queda em comparação a 2015 dado por $+8$.

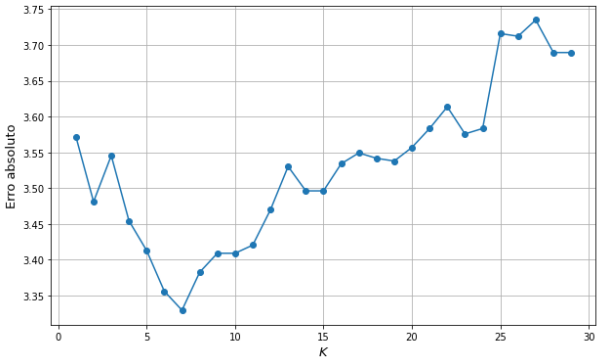
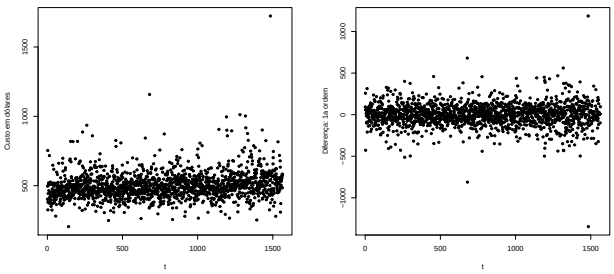


Figura 20: Gráfico da relação entre o número de vizinhos próximos e o erro absoluto médio associado à predição do índice político. Cada ponto representa a diferença absoluta entre o índice predito em cada um dos anos de 1950 a 2017 para K vizinhos próximos e seu respectivo valor exato.



(a) (b)

Figura 21: Série temporal original com o custo em dólares norte-americanos ao longo do tempo (a) e a diferença de primeira ordem (b) produzindo média estacionária.

Infelizmente não foi possível determinar correlações entre métricas econômicas e incidência de malária por conta da escassez de dados coletados. A coleta mundial, como mostrado na figura 16, considera apenas o período após 2000 que, quando avaliado o mesmo período em relação ao índice político mostrado na figura 17, não existe variação significativa a ponto de representar boas métricas

B. Séries temporais

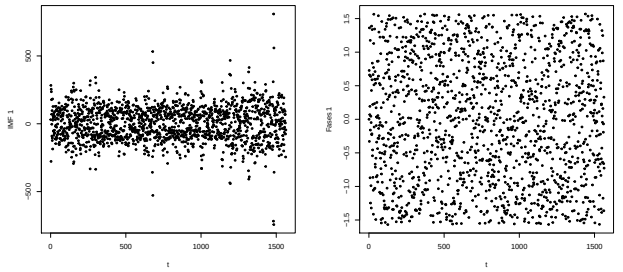
A análise da série temporal considerada [11] representa mais passos de processamento do que pré-processamento, diferentemente da análise para **IBL**.

Quando verificamos o gráfico relativo ao custo de analgésicos ao longo do tempo, notamos uma sutil (porém não desprezível) tendência crescente da média, como notamos na figura 21 (a). Para que possamos estudar a predição deste sistema, nossa primeira etapa é considerar a diferença de primeira ordem de cada ponto para que a série possa ser considerada estacionária. A nova série é mostrada na figura 21 (b)

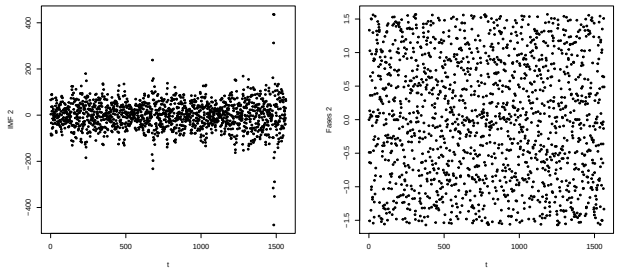
Com a nova série, podemos decompô-la em um determinado número de **IMF**s para delimitar um número de componentes predominantemente estocásticas e predominantemente determinísticas. Os resultados de cada uma das 10 **IMF**s encontradas podem ser observados nas figuras 22, 23, e 24. Além desta informação, nestas figuras decidimos representar a congruência de fase associada a cada componente de **IMF** para que quantitativamente possamos comparar este ”grau”de determinismo e estocasticidade e perceber mais evidentemente que decomposições de maiores graus condensam informações mais determinísticas. Em 24 (e) e (f) temos o resíduo da decomposição.

Para cada congruência de fase encontrada pelas 10 decomposições podemos então calcular a informação mútua **MI**, como representado na figura 25 (a). Neste momento, definimos um limiar $\epsilon = 0.01$ a partir do qual a **MI** será capaz de dividir em duas partes as 10 **IMF**s encontradas. Quando estabelecemos este valor, notamos que serão consideradas as 5 primeiras como componentes estocásticas e as 5 subsequentes como determinísticas. Esta $\mathbf{IMF}_5$ pode ser observada na figura 23 (a).

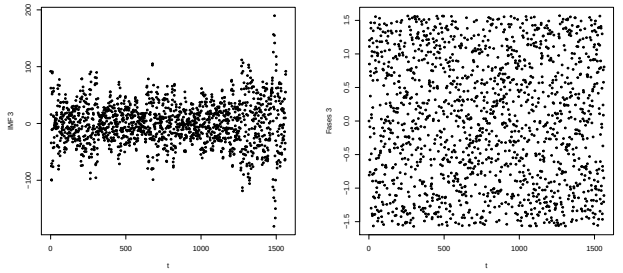
Com esta informação, para a modelagem das componentes determinísticas precisamos avaliar a informação mútua média **AMI** a elas relacionadas e o número de falsos vizinhos **FNN** encontrados. Para um atraso máximo de



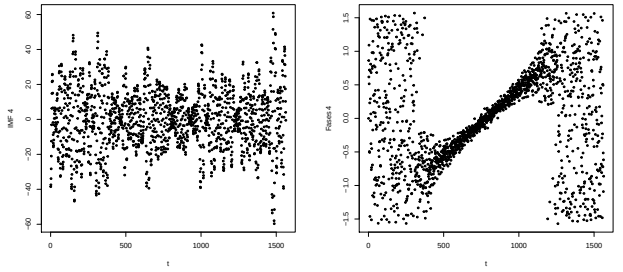
(a) (b)



(c) (d)

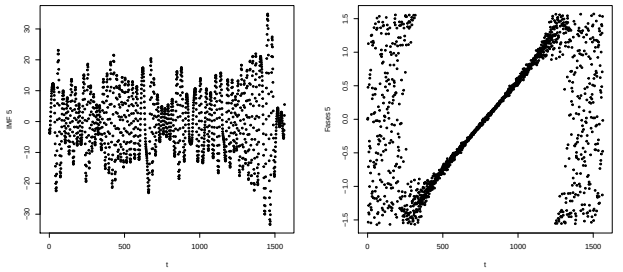


(e) (f)

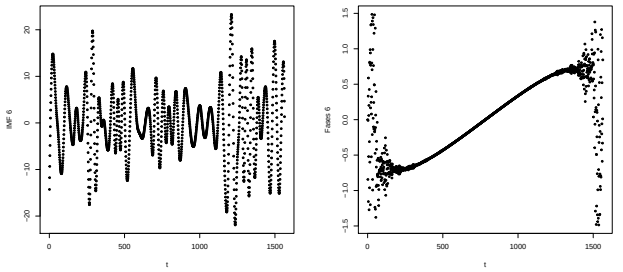


(g) (h)

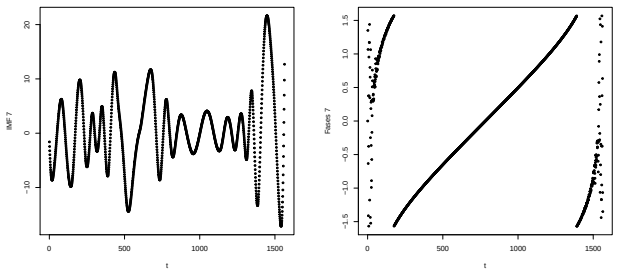
Figura 22: Decomposição em múltiplas **IMF**s para a série original. Apresentamos nas letras (a), (c), (e), (g) as **IMF**s de número 1 a 4, e nas letras (b), (d), (f), (h) a congruência de fase respectiva a cada **IMF** de número 1 a 4, respectivamente.



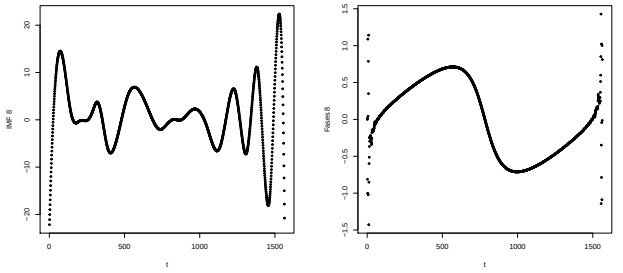
(a) (b)



(c) (d)

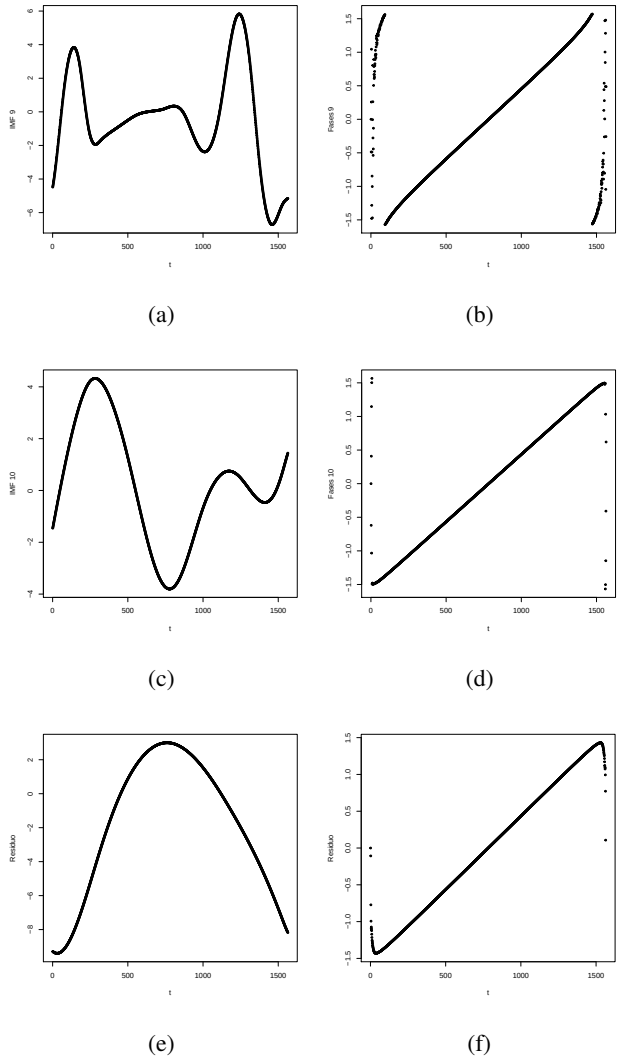


(e) (f)



(g) (h)

Figura 23: Decomposição em múltiplas **IMF**s para a série original. Apresentamos nas letras (a), (c), (e), (g) as **IMF**s de número 5 a 8, e nas letras (b), (d), (f), (h) a congruência de fase respectiva a cada **IMF** de número 5 a 8, respectivamente.



(a) (b)

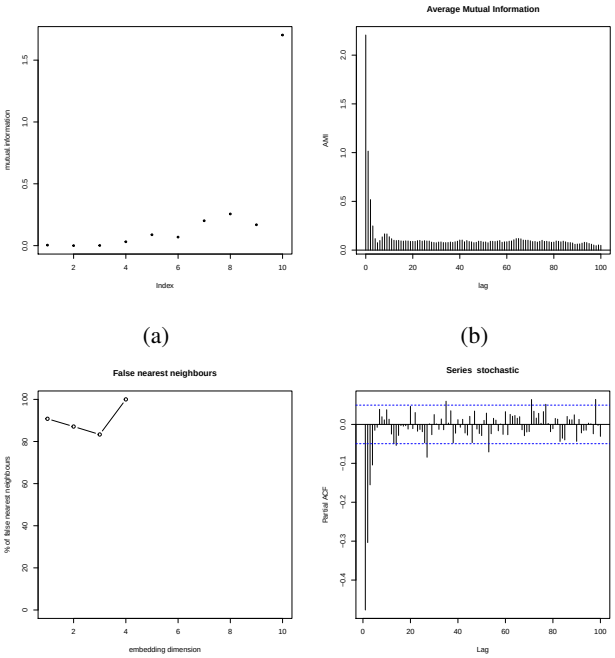
(c) (d)

(e) (f)

Figura 24: Decomposição em múltiplas **IMF**s para a série original. Apresentamos nas letras (a), (c), (e), as **IMF**s de número 9, 10 e resíduo, e nas letras (b), (d), (f) a congruência de fase respectiva a cada **IMF** de número 9, 10 e resíduo, respectivamente.

$lag = 100$, os resultados da **AMI** são mostrados na figura 25 (b) e, para um número máximo de *embedding dimension* $= 10$, apresentamos os resultados de **FNN** na figura 25 (c). Para nosso modelo determinístico obtivemos, assim, os parâmetros $d = 6$ e $m = 3$.

Infelizmente como notamos na figura 25 (c), o número de falsos vizinhos é bastante grande ($\approx 80\%$), mesmo no ponto de mínimo com $m = 3$. Isto prejudica a predição, uma vez que o número de dimensões pode não ser corretamente representado. Sabemos que o método **FNN** pode falhar na presença de muito ruído na série original [23]. Não foi possível descobrir se existe algum erro de implementação ou se alguma consideração sobre o sistema em análise foi precariamente feito e seguiremos a tentativa de reconstrução da série e subsequente predição de seu comportamento.



(a) (b)

(c) (d)

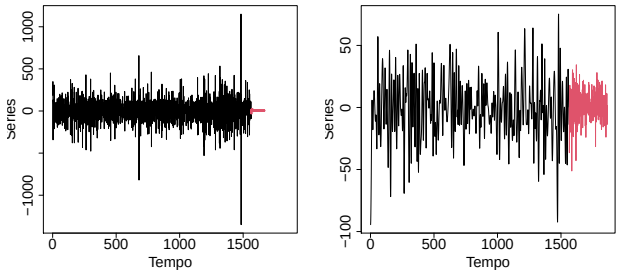
Figura 25: Resultados obtidos utilizando os algoritmos **MI** (a), para as 10 **IMF**s encontradas, **AMI** (b), para um máximo de $d = 100$ atrasos temporais para as componentes determinísticas; **FNN** (c), para um máximo de $m = 10$ dimensões para as **IMF**s determinísticas; e **PACF** (d), com atraso máximo de $d = 100$ para as componenetes estocásticas.

Por fim, para a componente estocástica, obtivemos por meio dos resultados da função **PACF** (figura 25 (d)) que o atraso temporal que otimiza a reconstrução do processo estocástico é dado por $d = 5$. Por outro lado, para obtenção do modelo completo, geramos 15 modelos

ARIMA com número de componentes determinísticos **AR** (Auto-Regressivos) variantes de 4 a 6 (dado que $d = 5$) e número de componentes estocásticos **MA** (Média Móvel) variantes de 0 a 14. Para cada modelo, obtivemos um valor de *Akaike Information Criterion* (**AIC**, ou Critério de Informação de Akaike). O modelo capaz de minimizá-lo [15] pode então ser finalmente somado ao modelo para as componentes determinísticas.

Na figura 26 apresentamos as componentes estocástica (a), determinística (b), com linhas em vermelho representando pontos preditos, e na figura (c), temos a série original (com diferença de ordem 1 para manter a tendência da média em zero) com valores preditos.

Na figura 27 optamos por representar medidas realizadas a partir do instante $t = 1300$ para evidenciar a zona predita. A predição efetiva é representada por uma linha contínua em preto dentro região em coloração roxa. A intensidade desta região representa o intervalo de confiança. Assim, a região com coloração mais intensa representa $\approx 68\%$ de acurácia.



(a) (b)

Figura 26: Predição do conjunto de componentes estocásticas (a) e determinísticas (b) ao longo do tempo. Em preto temos as linhas condizentes com os valores medidos e em vermelho, valores preditos.

Notamos que a predição apresenta tendência a média bastante evidente. Apesar disso, nossa modelagem para as componentes estocásticas (visto na figura 26 (a)) não se

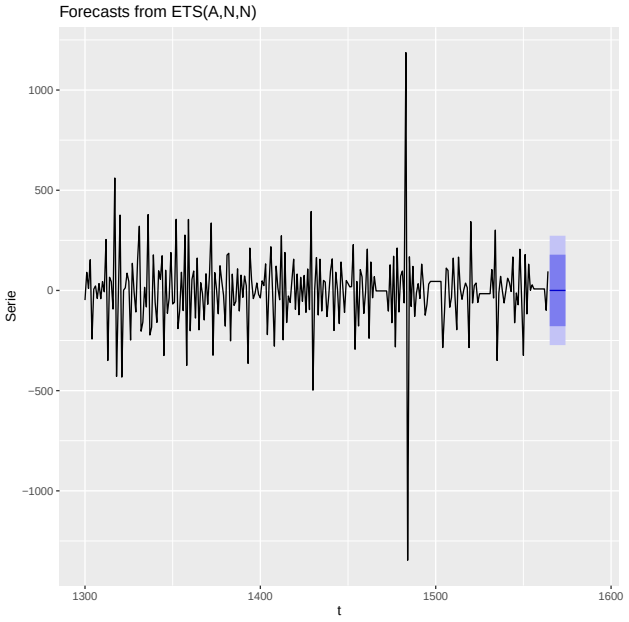


Figura 27: Representação da série original (com diferença de primeira ordem) en linha preta contínua e valores preditos na zona colorida em roxo. A coloração mais intensa representa o intervalo com $\approx 68.3\%$ de acurácia e a coloração mais clara, o intervalo cujos valores podem ser medidos no futuro com acurácia de $\approx 95.5\%$.

mostrou visualmente muito acurada. Isto porque conforme o tempo de predição avança, temos um rápido retorno à média em zero. Já para o modelo determinístico, visualmente temos um comportamento similar às medidas para tempos anteriores ao intervalo predito, apesar do número de **FNN** ter sido alto.

Neste sentido, é possível que outros métodos estocásticos possam ser mais adequados para o estudo desta série temporal, apesar da predição ter sido bem sucedida para tempos não muito distantes da última medida realizada.

VI. Conclusão

Com os resultados obtidos ao longo destes dois cenários mais abrangentes (**IBL** e Séries Temporais), pudemos conhecer melhor os modelos, algoritmos e as bases ma-

temáticas teóricas que os embasam. Os resultados obtidos para ambas não são discrepantes quando comparamos com aquilo que imaginávamos a princípio. Por exemplo para o cenário de regimes políticos, contamos que não existirão alterações bruscas no regime uma vez que experienciamos um período fortemente democrático nos últimos anos. O índice mantém-se positivo, mas ainda assim demonstra uma queda em sua pontuação.

Por outro lado a análise da série temporal nos indica que possivelmente o gasto manterá sua tendência de crescimento próximo à média. O resultado obtido aponta exatamente para este contexto como o mais provável nas predições.

Apesar desta coerência, é importante mencionar que muitos são os modelos em estudo para aprendizado de máquina, tanto em modelos baseados em instâncias quanto para modelos regressivos determinísticos quanto estocásticos. Esta pesquisa crescente deve ser vista por nós como motivadora para que sempre exista desenvolvimento de ciência básica e aplicada para que mais e melhores estudos teóricos e técnicos possam ser construídos e utilizados.

Referências

[1] Tatiane de Oliveira Silva Alencar. ≪Programa Farmacia Popular do Brasil: uma analise politica de sua origem, seus desdobramentos e inflexao≫. pt. Em: *Saude em Debate* 42 (out. de 2018), pp. 159–172. ISSN: 0103-1104. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-11042018000600159&nrm=iso.

[2] Yoann Bertrand, Karima Boudaoud e Michel Riveill. ≪What Do You Think About Your Company's Leaks? A Survey on End-Users Perception Toward Data Leakage Mechanisms≫. Em: *Frontiers in Big Data* 3 (2020), p. 38. ISSN: 2624-909X. DOI: 10.3389/fdata.2020.568257. URL: https://www.frontiersin.org/article/10.3389/fdata.2020.568257.

[3] Helga Malmin Binningsbø. ≪Power sharing, peace and democracy: Any obvious relationships?≫ Em: *International Area Studies Review* 16.1 (2013), pp. 89–112. DOI: 10.1177/2233865912473847. eprint: https://doi.org/10.1177/2233865912473847. URL: https://doi.org/10.1177/2233865912473847.

[4] Longbing Cao. ≪Data Science: A Comprehensive Overview≫. Em: *ACM Comput. Surv.* 50.3 (jun. de 2017). ISSN: 0360-0300. DOI: 10.1145/3076253. URL: https://doi.org/10.1145/3076253.

[5] Andrew M. Fraser e Harry L. Swinney. ≪Independent coordinates for strange attractors from mutual information≫. Em: *Phys. Rev. A* 33 (2 fev. de 1986), pp. 1134–1140. DOI: 10.1103/PhysRevA.33.1134. URL: https://link.aps.org/doi/10.1103/PhysRevA.33.1134.

[6] Francisco A. Gallego. ≪Historical Origins of Schooling: The Role of Democracy and Political Decentralization≫. Em: *The Review of Economics and Statistics* 92.2 (2010), pp. 228–243. DOI: 10.1162/rest.2010.11894. eprint: https://doi.org/10.1162/rest.2010.11894. URL: https://doi.org/10.1162/rest.2010.11894.

[7] Gwilym M. George Box e Gregory Reinsel Jenkins. *Time series analysis: Forecasting and control*. 3rd. Prentice Hall, 1994.

[8] Clive W.J. Granger. ≪Time Series Analysis, Cointegration, and Applications≫. Em: *American Economic Review* 94.3 (jun. de 2004), pp. 421–425. DOI: 10.1257/0002828041464669. URL: https://www.aeaweb.org/articles?id=10.1257/0002828041464669.

[9] Norden E. Huang e Zheng Shen. ≪The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis≫.

Em: *Proceedings of the Royal Society A* 454 (1971 1998), pp. 903–995.

[10] Norden E. Huang, Man-Li Wu e Wendong Qu. ≪Applications of Hilbert–Huang transform to non-stationary financial time series analysis≫. Em: *Applied Stochastic Models in Business and Industry* 19.3 (2003), pp. 245–268. DOI: https://doi.org/10.1002/asmb.501. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asmb.501. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.501.

[11] Shruti Kaushik et al. ≪AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures≫. Em: *Frontiers in Big Data* 3 (2020), p. 4. ISSN: 2624-909X. DOI: 10.3389/fdata.2020.00004. URL: https://www.frontiersin.org/article/10.3389/fdata.2020.00004.

[12] Matthew B. Kennel, Reggie Brown e Henry D. I. Abarbanel. ≪Determining embedding dimension for phase-space reconstruction using a geometrical construction≫. Em: *Phys. Rev. A* 45 (6 mar. de 1992), pp. 3403–3411. DOI: 10.1103/PhysRevA.45.3403. URL: https://link.aps.org/doi/10.1103/PhysRevA.45.3403.

[13] M. G. Marshall e Elzinga-Marshall G. C. *Global Report 2017 - Conflict, Governance and State Fragility*. Center for Systemic Peace, 2017, pp. 29–31.

[14] Joseph Frank Mgaya. ≪Application of ARIMA models in forecasting livestock products consumption in Tanzania≫. Em: *Cogent Food & Agriculture* 5.1 (2019). Ed. por Fatih Yildiz, p. 1607430. DOI: 10.1080/23311932.2019.1607430.

[15] ≪Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions≫. Em: *Psychometrika* 52 (3 1987), pp. 345–370.

[16] George Nasser, Ben W. Morrison e Piers Bayl-Smith. ≪The Role of Cue Utilization and Cognitive Load in the Recognition of Phishing Emails≫. Em: *Frontiers in Big Data* 3 (2020), p. 33. ISSN: 2624-909X. DOI: 10.3389/fdata.2020.546860. URL: https://www.frontiersin.org/article/10.3389/fdata.2020.546860.

[17] OECD. *OECD - Gender Equality*. 2020. URL: https://www.oecd.org/gender/data/.

[18] N. H. Packard e J. P. Crutchfield. ≪Geometry from a Time Series≫. Em: *Phys. Rev. L* 45 (9 nov. de 1980), pp. 712–716.

[19] Ben Palmquist. ≪Equity, Participation, and Power: Achieving Health Justice Through Deep Democracy≫. Em: *The Journal of Law, Medicine & Ethics* 48.3 (2020). PMID: 33021188, pp. 393–410. DOI: 10.1177/1073110520958863. eprint: https://doi.org/10.1177/1073110520958863. URL: https://doi.org/10.1177/1073110520958863.

[20] *Political Conflict Data - Political Science - Research Guides at New York University*. URL: https://guides.nyu.edu/polisci/political-conflict-data.

[21] Hongchao Qi e Shuang Xiao. ≪COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis≫. Em: *Science of The Total Environment* 728 (2020), p. 138778. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2020.138778. URL: http://www.sciencedirect.com/science/article/pii/S0048969720322956.

[22] Dan Reiter. ≪Does Peace Nurture Democracy?≫ Em: *The Journal of Politics* 63.3 (2001), pp. 935–948. DOI: 10.1111/0022-3816.00095. eprint: https://doi.org/10.1111/0022-3816.00095. URL: https://doi.org/10.1111/0022-3816.00095.

[23] Carl Rhodes e Manfred Morari. ≪The false nearest neighbors algorithm: An overview≫. Em: *Computers & Chemical Engineering* 21 (1997). Supplement to Computers and Chemical Engineering,

S1149–S1154. ISSN: 0098-1354. DOI: https://doi.org/10.1016/S0098-1354(97)87657-0. URL: http://www.sciencedirect.com/science/article/pii/S0098135497876570.

[24] Ricardo Araújo Rios. ≪Improving time series modeling by decomposing and analysing stochastic and deterministic influences≫. Tese de doutoramento. Universidade de São Paulo.

[25] Ricardo Araújo Rios e Rodrigo Fernandes de Mello. ≪Applying Empirical Mode Decomposition and mutual information to separate stochastic and deterministic influences embedded in signals≫. Em: *Signal Processing* 118 (2016), pp. 159–176. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2015.07.003. URL: http://www.sciencedirect.com/science/article/pii/S0165168415002297.

[26] Max Roser. *'Malaria' Our World in Data*. 2016. URL: https://www2.deloitte.com/us/en/insights/industry/health-care/forces-of-change-health-care.html.

[27] Max Roser. *Democracy - Our World in Data*. URL: https://ourworldindata.org/democracy.

[28] C. E. Shannon. ≪A Mathematical Theory of Communication≫. Em: *The Bell System Technical Journal* 27 (out. de 1948), pp. 379–423.

[29] Tânia Tomé e Mário J. de Oliveira. *Stochastic Dynamics and Irreversibility*. 2015. ISBN: 978-3-319-11769-0. DOI: 10.1007/978-3-319-11770-6. URL: http://link.springer.com/10.1007/978-3-319-11770-6.

[30] A. Witt, J. Kurths e A. Pikovsky. ≪Testing stationarity in time series≫. Em: *Phys. Rev. E* 58 (2 ago. de 1998), pp. 1800–1810. DOI: 10.1103/PhysRevE.58.1800. URL: https://link.aps.org/doi/10.1103/PhysRevE.58.1800.

[31] Vikram Kumar Yeragani e K.A.Radha Krishna Rao. ≪Diminished chaos of heart rate time series in patients with major depression≫. Em: *Biological Psychiatry* 51.9 (2002), pp. 733–744. ISSN: 0006-3223. DOI: https://doi.org/10.1016/S0006-3223(01)01347-6. URL: http://www.sciencedirect.com/science/article/pii/S0006322301013476.