UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS JURÍDICAS E ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

ALBERSON DA SILVA MIRANDA

MÉTODOS DE MACHINE LEARNING PARA RECONCILIAÇÃO ÓTIMA DE SÉRIES TEMPORAIS HIERÁRQUICAS E AGRUPADAS

VITÓRIA

2023

ALBERSON DA SILVA MIRANDA

MÉTODOS DE MACHINE LEARNING PARA RECONCILIAÇÃO ÓTIMA DE SÉRIES TEMPORAIS HIERÁRQUICAS E AGRUPADAS

Dissertação apresentada ao Programa de Pós-Graduação em Economia da Universidade Federal do Espírito Santo, como requisito para a obtenção do título de Mestre em Economia.

Orientador: Prof. Dr. Guilherme A. A. Pereira

VITÓRIA
2023

ALBERSON DA SILVA MIRANDA
MÉTODOS DE MACHINE LEARNING PARA RECONCILIAÇÃO ÓTIMA DE SÉRIES TEMPORAIS HIERÁRQUICAS E AGRUPADAS/ ALBERSON DA SILVA MIRANDA. – VITÓRIA, 2023-
41p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Guilherme A. A. Pereira

Dissertação (Mestrado) – UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS JURÍDICAS E ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA, 2023.

1. Palavra-chave1. 2. Palavra-chave2. 2. Palavra-chave3. I. Orientador. II. Universidade xxx. III. Faculdade de xxx. IV. Título

ALBERSON DA SILVA MIRANDA

MÉTODOS DE MACHINE LEARNING PARA RECONCILIAÇÃO ÓTIMA DE SÉRIES TEMPORAIS HIERÁRQUICAS E AGRUPADAS

Dissertação apresentada ao Programa de Pós-Graduação em Economia da Universidade Federal do Espírito Santo, como requisito para a obtenção do título de Mestre em Economia.

Aprovada em xx de xx de 20xx.

COMISSÃO EXAMINADORA

Prof. Dr. Guilherme A. A. Pereira
Universidade Federal do Espírito Santo
Orientador

Professor
Instituição

Professor
Instituição

RESUMO

Sed mattis, erat sit amet gravida malesuada, elit augue egestas diam, tempus scelerisque nunc nisl vitae libero. Sed consequat feugiat massa. Nunc porta, eros in eleifend varius, erat leo rutrum dui, non convallis lectus orci ut nibh. Sed lorem massa, nonummy quis, egestas id, condimentum at, nisl. Maecenas at nibh. Aliquam et augue at nunc pellentesque ullamcorper. Duis nisl nibh, laoreet suscipit, convallis ut, rutrum id, enim. Phasellus odio. Nulla nulla elit, molestie non, scelerisque at, vestibulum eu, nulla. Ut odio nisl, facilisis id, mollis et, scelerisque nec, enim. Aenean sem leo, pellentesque sit amet, scelerisque sit amet, vehicula pellentesque, sapien.

Palavras-chave: palavra-chave1. palavra-chave2. palavra-chave3.

ABSTRACT

Sed mattis, erat sit amet gravida malesuada, elit augue egestas diam, tempus scelerisque nunc nisl vitae libero. Sed consequat feugiat massa. Nunc porta, eros in eleifend varius, erat leo rutrum dui, non convallis lectus orci ut nibh. Sed lorem massa, nonummy quis, egestas id, condimentum at, nisl. Maecenas at nibh. Aliquam et augue at nunc pellentesque ullamcorper. Duis nisl nibh, laoreet suscipit, convallis ut, rutrum id, enim. Phasellus odio. Nulla nulla elit, molestie non, scelerisque at, vestibulum eu, nulla. Ut odio nisl, facilisis id, mollis et, scelerisque nec, enim. Aenean sem leo, pellentesque sit amet, scelerisque sit amet, vehicula pellentesque, sapien.

Keywords: keyword1. keyword2. keyword3.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MinT	*Minimum Trace*
MCRL	Modelo Clássico de Regressão Linear
MQO	Mínimos Quadrados Ordinários
MQP	Mínimos Quadrados Ponderados
ANN	*Artificial Neural Network*
SVR	*Support Vector Regression*
SFN	Sistema Financeiro Nacional
FAVAR	*Factor Augmented Vector Autoregression*

LISTA DE SÍMBOLOS

t	Tempo dentro da amostra
T	Último tempo dentro da amostra, quantidade de observações numa série
h	Horizonte de previsão, tempo fora da amostra
Ω	Conjunto de dados dentro da amostra
y	Série temporal dentro da amostra
$\hat{y}$	Série temporal estimada
$\tilde{y}$	Série temporal reconciliada
n	Número de séries na hierarquia
m	Número de séries no menor nível da hierarquia
k	Número de níveis na hierarquia
$\boldsymbol{S}$	Matriz de soma
$\boldsymbol{G}$	Matriz de reconciliação
$\{...\}$	Conjunto
$\lvert\{...\}\rvert$	Cardinalidade de um conjunto

SUMÁRIO

1 INTRODUÇÃO

[Escrever *outline* da dissertação]

1.1 Motivação

Embora no séc. XX ainda houvesse espaço para uma gestão guiada apenas por instinto (WALLANDER, 1999), atualmente é impensável um banco não realizar previsões de seus resultados e comunicar suas expectativas ao mercado. Nesse documento, ou *guidance*, a projeção da carteira de crédito — o total de empréstimos e financiamentos, dentre outros itens — é frequentemente a primeira informação fornecida, uma vez que é um dos principais elementos para o planejamento dos bancos comerciais. Juntamente com as projeções de depósitos, provisões para créditos de liquidação duvidosa, eficiência operacional, entre outros indicadores-chave, essas projeções determinam a temperatura das expectativas da instituição em relação a elementos cruciais como rentabilidade, dividendos e posição no mercado (*market-share*), e isso é essencial para os acionistas e investidores. Essas projeções precisam ser tão precisas quanto possível, para que se possa calcular o risco de transacionar com a instituição financeira, seja como investidor ou cliente.

Ainda que não exista penalidades específicas para instituições financeiras que erram (por uma boa margem) em suas projeções, elas podem sofrer consequências negativas em outros aspectos, como na avaliação de seus desempenhos por parte dos investidores e clientes. Os investidores e clientes podem considerar as projeções equivocadas como um sinal de falta de competência ou confiança na instituição financeira, o que pode afetar negativamente a reputação e a imagem da instituição. Isso pode levar a uma redução no número de investimentos e depósitos, o que irá afetar diretamente sua saúde financeira.

Além disso, nos casos em que algum grupo se sentir lesado, as instituições financeiras podem enfrentar ações judiciais se suas projeções forem consideradas enganosas ou fraudulentas. Por exemplo, se uma instituição financeira fizer projeções excessivamente otimistas para incentivar os investidores a comprar seus títulos e, posteriormente, as projeções se mostrarem incorretas, ela pode ser acusada de fraude[1] ou, ao menos, gestão temerária[2] — ambos caracterizados como crime contra o Sistema Financeiro Nacional.

Não são raros os casos em que bancos manipulam seus demonstrativos para se apresentarem mais saudáveis do que realmente o são para atrair clientes e investidores. Em 2001, o Banco Santos era classificado como o oitavo maior banco brasileiro. Em 2004 se encontrava sob intervenção do Bacen: ao mesmo tempo em que continuava a expandir seus negócios, o banco

[1] Art. 3º: Divulgar informação falsa ou prejudicialmente incompleta sobre instituição financeira. Pena: Reclusão, de 2 (dois) a 6 (seis) anos, e multa. Art. 4º: Gerir fraudulentamente instituição financeira. Pena: Reclusão, de 3 (três) a 12 (doze) anos, e multa (BRASIL, 1986).

[2] Art. 4º, parágrafo único: Se a gestão é temerária: Pena: Reclusão, de 2 (dois) a 8 (oito) anos, e multa (BRASIL, 1986).

escondia um rombo de cerca de R$ 2,2 bi em créditos que já haviam sido liquidado mas eram mantidos no balanço (MOURA, 2007). De forma semelhante, em 2009, o Banco PanAmericano publicou em suas demonstrações contábeis carteira de crédito de R$ 9,9 bi. Entretanto, cerca de 25% já havia sido liquidado e era mantido no balanço artificialmente como objetivo de evitar a redução da atratividade do banco no mercado (COELHO et al., 2015).

Por isso, é importante que as instituições financeiras sejam transparentes e precisas em suas projeções, fornecendo informações confiáveis e atualizadas para seus clientes e investidores. No entanto, há também motivações estratégicas para essa atividade. Beccalli et al. (2015) mostraram que, em uma amostra de 55 bancos europeus, a utilização de *guidance* está associada a um aumento de 15% na probabilidade do banco atingir ou superar as expectativas de mercado. Isso, por sua vez, está associado a um incremento de até 5% no retorno por ação em relação aos bancos que não alcançaram ou superaram as expectativas.

No que concerne a elaboração dessas previsões, a prática usual em *budgeting*, principalmente para empresas com muitas filiais, é a *top-down*, ou seja, realizar previsões para o agregado e então distribuí-las para os níveis desagregados seguindo algum método para desagregação. No caso dos bancos de varejo, com muitas agências espalhadas pelo território, especialmente em um país grande como o Brasil, esse método é muito prático.

Esse é o caso do Banestes. Com 96 agências distribuídas pelos 78 municípios capixabas, realizar o *budgeting* para R$ 5,5 bi de faturamento não é uma tarefa trivial. Além de uma estrutura hierárquica larga, se tratando de um banco múltiplo[3] que opera com diversas carteiras, as n modalidades de crédito[4] expandem a estrutura para um total de $n \times 96$ séries temporais a serem estimadas.

Dada tal complexidade, a abordagem *top-down* se coloca como uma opção viável em termos de tempo de processamento e análise, principalmente nos maiores níveis de agregação. No entanto, conforme descemos na hierarquia, menos precisa ela se torna e, além disso, as características individuais das séries temporais do menor nível hierárquicos são ignoradas. Isso significa que, se no agregado a previsão para uma carteira de crédito for de crescimento de 10%, todas as 96 agências devem seguir a mesma estimativa, divergindo apenas na proporção de participação de cada uma no total.

Tomando o caminho inverso, a abordagem *bottom-up* consiste em realizar previsões para cada série temporal individualmente e, então, agregá-las para obter a previsão para o total. Essa abordagem pode ser mais precisa, pois leva em consideração as características individuais de cada série temporal do nível mais desagregado. No entanto, ela é mais custosa em termos de tempo de processamento e análise. Nesse sentido, cabe ao analista avaliar o *trade-off* entre os

[3] Para ser classificado como banco múltiplo, a instituição financeira deve operar com, no mínimo, duas carteiras dentre: comercial; investimento ou desenvolvimento; crédito imobiliário; de crédito, financiamento e investimento, e; arrendamento mercantil (CONSELHO MONETÁRIO NACIONAL, 1994).

[4] Crédito consignado, rural, imobiliário, pessoal, capital de giro, desconto de títulos etc.

ganhos de precisão percebidos com a geração de previsões individuais e a economia de tempo e processamento em realizar o contrário (GROSS; SOHL, 1990).

Além disso, ambas são abordagens de nível único, isto é, são realizadas as previsões para um único nível e então os demais níveis são obtidos agregando ou desagregando. O problema com esses tipos de abordagem é que elas utilizam informação imcompleta (HYNDMAN; ATHANASOPOULOS, 2021). Por exemplo, suponha-se que se escolha estimar modelos para cada uma das 96 agências e agregá-las (*bottom-up*). Nesse caso, ignora-se a influência que os níveis mais agregados — aqui a carteira de crédito da região ou de todo o estado — pode ter na estimação do saldo de crédito de cada agência. Por outro lado, se escolher estimar modelos para os níveis mais agregados (*top-down*), ignora-se a informação individual de cada agência.

A reconciliação ótima de previsões pontuais é uma abordagem que busca resolver esse problema. Ela consiste em realizar previsões para todos os níveis hierárquicos e, então, estimar um modelo para reescrever as previsões do nível mais desagregado como uma combinação linear de todos os elementos da hierarquia, para então agregá-las de forma semelhante ao *bottom-up*, obtendo previsões coerentes nos níveis superiores. Dessa forma, a informação de todos os níveis é utilizada na estimação dos modelos e na geração das previsões, ao mesmo tempo em que a variância do erro de previsão é minimizado (HYNDMAN; AHMED et al., 2011).

Atualmente, os métodos analíticos, especificamente o *Minimum Trace* (WICKRAMASURIYA; ATHANASOPOULOS; HYNDMAN, 2019), são os mais populares na literatura da reconciliação ótima. Entretanto, tais métodos são sujeitos a uma série de restrições, como as do MCLR, e têm sua capacidade preditiva reduzida quando suas hipóteses são violadas.

Em previsões de séries temporais, o objetivo na maioria dos casos é prever valores futuros com a maior acurácia possível. Em vista disso, métodos de *machine learning* são mais gerais, no sentido de permitir parâmetros não lineares e poderem aproximar virtualmente qualquer função. Além disso, são focados na capacidade preditiva, muitas vezes em detrimento da explicativa. Espera-se, portanto, que esses métodos alcancem melhor performance no problema da reconciliação ótima, justificando a pesquisa e atenção ao tema.

1.2 Objetivos

O objetivo geral da dissertação é estudar o problema da reconciliação ótima de previsões pontuais a partir de métodos de *machine learning*.

Como objetivos específicos, tenho:

1. Estudar métodos para estimação da matriz de reconciliação aplicando algoritmos e fluxos de trabalho de *machine learning*, como *tuning* e *resampling*;
2. Identificar possíveis vantagens e limitações da abordagem por *machine learning* na reconciliação de previsões pontuais a partir de aplicação do métodos estudados na previsão de

saldos de crédito do Banestes.

2 REVISÃO DE LITERATURA

2.1 Previsão de saldos de crédito de instituições financeiras

A nível macroeconômico, a previsão do agregado de crédito das instituições financeiras é uma preocupação de bancos centrais ao redor do mundo. No Brasil, Bader, Koyama, Sérgio Mikio e Tsuchida, Marcos Hiroyuki (2014) aprimoram o método FAVAR com uma etapa de análise de correlação canônica para identificar as melhores, em termos de correlação com as variáveis de crédito do SFN, combinações lineares de componentes principais. Esse método, que chamaram de FAVAR canônico, alcançou resultado superior aos FAVAR em 1 e 2 estágios na previsão das variáveis de crédito utilizadas, que foram: a concessão de crédito total com recursos livres, o saldo da carteira de crédito total com recursos livres, o saldo da carteira de crédito total com recursos direcionados, a taxa de inadimplência da carteira de crédito total com recursos livres e a taxa média de juros das operações de crédito total com recursos Livre. O trabalho abordou apenas o nível mais agregado, no total do SFN.

Çolak et al. (2019) produzem uma série de indicadores para monitoramento dos períodos de expansão e desaceleração moderada ou excessiva de crédito no setor bancário turco. Os autores utilizam séries filtradas do agregados de crédito, crédito comercial, crédito direto ao consumidor, crédito imobiliário e fincanciamento de veículos, além de diversos setores da economia, como agricultura, manufatura, construção, comércio, dentre outros, para prever os ciclos de crédito no sistema bancário turco. Os autores concluem que os indicadores com maior poder de explicação para as variáveis macroeconômicas são a taxa de crescimento real do crédito e a taxa de resposta ao impulso do crédito.

Já para niveis abaixo do agregado de crédito, poucos trabalhos foram encontrados. Tangenciando o tema da previsão de saldos de crédito, outros tópicos da economia bancária foram objeto de estudo para previsão de séries temporais. Sezer, Gudelek e Ozbayoglu (2019) produziram revisão de literatura de trabalhos publicados entre 2005 e 2019 que realizaram previsão de séries temporais financeiras utilizando *deep learning* e os agruparam em preços de ações individuais, índices (e.g., IBovespa, Dow Jones), preços de commodities, tendência e volatilidade de ativos, preços de títulos, câmbio e preços de criptomoedas. Apesar da extensa revisão, não foram encontrados trabalhos que combinassem estruturas hierárquicas com *machine learning*.

Gorodetskaya, Gobareva e Koroteev (2021) fornecem uma metodologia "universal" para aplicação automática de *machine learning* na previsão de séries temporais do setor bancário, que poderia ser aplicada em qualquer tipo de problema. A metodologia consiste em obter preditores a partir da própria variável defasada, das estatísticas básicas da variável (máximo, mínimo, média, variância etc.), e de anomalias periódicas detectadas, e selecioná-las pela medida de importância.

Realizaram uma revisão de literatura recente sobre o assunto e apresentaram sua abordagem para o problema da previsão da demanda por moeda em caixas eletrônicos.

No que diz respeito à previsão de séries temporais em largas hierarquias, Prayoga, Suhartono e Rahayu (2017) trabalharam na previsão do fluxo de caixa do Banco da Indonésia, utilizando uma hierarquia de 3 níveis — 40 agências no nível mais desagregado, as 6 grandes ilhas do país como nível intermediário e o total no nível mais agregado. Os autores realizaram um *benchmark* de 5 modelos para previsão da série no nível mais agregado e utilizando o método *top-down* para obter as previsões no nível mais desagregado, concluindo pela efetividade do método *top-down* por proporções históricas. Entretanto, os autores não incluíram reconciliação ótima, a estimativa *bottom-up* ou mesmo outros métodos *top-down* para efeito de comparação, o que limita o alcance do trabalho.

2.2 Previsão de séries temporais hierárquicas e agrupadas

A segunda etapa da revisão de literatura consistiu na pesquisa bibliográfica relacionada à reconciliação ótima de previsões de séries temporais hierárquicas e agrupadas e sua interseção com o tema *machine learning*.

2.2.1 Abordagens de nível único

Uma abordagem de nível único é uma abordagem em que as previsões são realizadas para um único nível da hierarquia. A partir dessas previsões, os demais níveis são obtidos, ou desagregando (no caso dos níveis inferiores), ou agregando (no caso dos níveis superiores) essas informações (HYNDMAN; ATHANASOPOULOS, 2021). Os métodos *top-down*, *bottom-up* e *middle-out* são abordagens de nível único.

Enquanto há apenas uma única forma de se agregar níveis na hierarquia (*bottom-up*), a desagregação (*top-down*) pode ser realizada de, ao menos, duas dezenas de maneiras (GROSS; SOHL, 1990). Dois dos métodos mais intuitivos são a média das proporções históricas e a proporção das médias históricas.

Na média das proporções históricas, cada proporção p_j, com $j = 1, ..., m$, consiste em tomar a média das proporções da série desagregada $y_{j,t}$ em relação ao agregado y_t:

$$p_j = \frac{1}{T} \sum_{t=1}^{T} \frac{y_{j,t}}{y_t} \tag{1}$$

Já a proporção das médias históricas consiste em tomar a proporção das médias das séries desagregadas em relação à média do agregado[5].

[5] Isso é equivalente a tomar a proporção direta entre os somatórios das séries. Note que, pelas propriedades do operador de somatório, $\sum_{t=1}^{T} \frac{y_t}{T} = \frac{y_1}{T} + \ldots + \frac{y_T}{T} = \frac{y_1 + \ldots + y_T}{T} = \frac{\sum_{t=1}^{T} y_t}{T}$. Então, a equação Equação 2

$$p_j = \frac{\sum_{t=1}^{T} \frac{y_{j,t}}{T}}{\sum_{t=1}^{T} \frac{y_t}{T}} \tag{2}$$

Athanasopoulos, Ahmed e Hyndman (2009) desenvolvem o método proporções de previsão, que consiste em um método *top-down* em que os pesos são calculados a partir das proporções das previsões base ao invés do passado. A vantagem do método é que os pesos estarão os mais próximos das características mais recentes da série, ao invés de serem baseados em dados históricos. A desvantagem é que se deve realizar previsões para toda a hierarquia, perdendo o ganho de agilidade dos demais métodos *top-down*.

$$p_j = \prod_{i=0}^{K-1} \frac{\hat{Y}_{j,n}^{(i)}(h)}{\sum(\hat{Y}_{j,n}^{(i+1)}(h))}$$

Li et al. (2016) compararam dois algoritmos de *machine learning* para previsão da produção de energia solar no estado da Flórida/EUA: ANN e SVR. Argumentando que tradicionalmente as previsões nesse problema são realizadas com os dados de produção total da planta, eles propõem uma abordagem hierárquica *bottom-up*, com previsões base de cada inversor solar. Os autores concluem que a abordagem hierárquica *bottom-up* é mais precisa do que a previsão do agregado, ao menos na previsão um passo a frente. Embora os autores utilizem algoritmos de *machine learning* para as previsões base, eles não utilizam esses algoritmos para a reconciliação ótima, caracterizando a abordagem do trabalho ainda como nível único.

2.2.2 Métodos analíticos para reconciliação ótima

Previsões pontuais de séries temporais hierárquicas não é um assunto novo. Ao menos desde a década de 70, pesquisas foram publicadas acerca de abordagens *bottom-up* e *top-down*, suas vantagens e desvantagens, e tentativas de se definir qual é o melhor método[6]. Entretanto, é apenas em Hyndman, Ahmed et al. (2011) que é formalizada uma abordagem prática que utiliza toda a informação disponível, (i.e. as previsões de todos elementos de todos os níveis da hierarquia) a partir da estimação da matriz $\boldsymbol{G}$ via regressão linear por mínimos quadrados generalizados (MQG).

Entretanto, para ser capaz de estimar o modelo por MQG, é necessária a matriz de variância-covariância dos erros. Hyndman, Ahmed et al. (2011) usam a matriz de erros de coerência, ou seja, a diferença entre as previsões reconciliadas e as previsões base, que tem posto incompleto e não identificada e, portanto, não pode ser estimada. Os autores contornam esse problema adotando no lugar da matriz de variância-covariância dos erros uma matriz diagonal

pode ser simplificada para $p_j = \frac{\sum_{t=1}^{T} y_{j,t}}{\sum_{t=1}^{T} y_t}$.

[6] Uma revisão dessa literatura pode ser encontrada em Athanasopoulos, Ahmed e Hyndman (2009).

constante, ou seja, assumem variância constante dos erros de reconciliação, e estimam a matriz $\boldsymbol{G}$ por mínimos quadrados ordinários (MQO).

A estimação por esse método resulta numa reconciliação ótima que depende apenas da matriz $\boldsymbol{S}$, ou seja, da estrutura hierárquica, e independe da variância e covariância das previsões base $\hat{\boldsymbol{y}}_{\boldsymbol{T+h}}$ — o que não é uma conclusão satisfatória.

Hyndman, Lee e Wang (2016) tentam aperfeiçoar o método usando as variâncias das previsões base estimadas (dentro da amostra) como estimativa para a matriz de variância-covariância dos erros de reconciliação, de forma a as utilizar como pesos e realizar a reconciliação ótima por mínimos quadrados ponderados (MQP). Assim, previsões base mais acuradas têm peso maior do que as mais ruidosas. Entretanto, não fornecem justificativa teórica para usar a diagonal da matriz de variância-covariância de $\hat{\boldsymbol{e}}_{\boldsymbol{t}}$.

Wickramasuriya, Athanasopoulos e Hyndman (2019) argumentam que o que de fato interessa é que as previsões reconciliadas tenham o menor erro. Então, corrigem a abordagem de reconciliação ótima para o objetivo de minimização dos erros das previsões reconciliadas $\tilde{\boldsymbol{y}}_{\boldsymbol{t+h}}$, ao invés dos erros das previsões base $\hat{\boldsymbol{y}}_{\boldsymbol{t+h}}$. Dado que isso implica na minimização da variância de $\tilde{\boldsymbol{e}}_{\boldsymbol{t+h}}$, ou seja, na minimização do somatório da diagonal, o traço, da matriz de variância-covariância de $\tilde{\boldsymbol{e}}_{\boldsymbol{t+h}}$, eles chamaram esse método de Traço Mínimo (MinT, na sigla em inglês). Paralelamente, usam desigualdade triangular para demonstrar que as previsões reconciliadas obtidas por esse método são ao menos tão boas quanto as previsões base.

Panagiotelis et al. (2021) reinterpreta a literatura de coerência e reconciliação de previsões pontuais a partir de uma abordagem geométrica, trazendo provas alternativas para conclusões anteriores ao mesmo tempo em que fornece novos teoremas. Além disso, os autores estendem essa interpretação geométrica para o contexto probabilístico, fornecendo métodos paramétricos e não paramétricos (via *bootstrapping*) para reconciliação de previsões probabilísticas, ou seja, para reconciliar previsões $\hat{y}_t$ obtidas a partir de toda a distribuição, e não apenas a média.

2.2.3 Métodos de machine learning para reconciliação ótima

Spiliotis et al. (2021) propõem a utilização de *machine learning* para a reconciliação ótima de séries temporais, especificamente os algoritmos de árvore de decisão *Random Forest* e *XGBoost*. Os autores descrevem como vantagens desse método em relação aos anteriores a descrição de relacionamentos não lineares, performance preditiva e a desnecessidade da utilização de todos os elementos da hierarquia na combinação ótima. A abordagem utilizada foi:

1. dividir a amostra em treino e teste;
2. treinar um modelo de previsão na amostra treino e obter previsões um passo a frente para a amostra teste;

3. treinar um modelo de *machine learning* para cada série do menor nível da hierarquia, em que os parâmetros são as previsões obtidas no passo 2 e a variável explicada são os valores observados. Isso resulta em um modelo de reconciliação ótima para cada elemento do menor nível da hierarquia, combinando informações disponíveis de todos os níveis hierárquicos;
4. obter as previsões base $\hat{y}_t$;
5. passar as previsões base ao modelo treinado no passo 3 para se obter as previsões reconciliadas para o menor nível da hierarquia;
6. agregar as previsões reconciliadas para se obter as previsões nos demais níveis hierárquicos.

Para o conjunto de dados utilizados, Spiliotis et al. (2021) afirmam que os métodos de *machine learning*, especialmente o XGBoost, alcançaram, em média, melhor performance que as abordagens de nível único e o MinT. Além disso, concluíram que quanto maior é a diferença entre as séries, em todos os níveis hierárquicos, maior são os benefícios da abordagem por *machine learning*.

3 MÉTODOS PARA RECONCILIAÇÃO DE SÉRIES TEMPORAIS

3.1 Séries hierárquicas e séries agrupadas

Séries temporais hierárquicas são aquelas que podem ser agregadas ou desagregadas naturalmente em uma estrutura aninhada (HYNDMAN; ATHANASOPOULOS, 2021). Para ilustrar, tome a série do PIB brasileiro. Ela pode ser desagregada por estado que, por sua vez, pode ser desagregada por município.

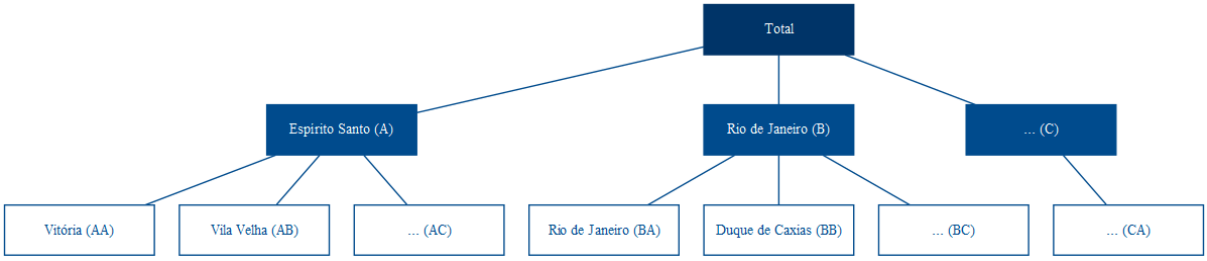

Figura 1 – Séries Hierárquicas

Essa estrutura pode ser representada por equações para qualquer nível de agregação.

$$y_t = y_{A,t} + y_{B,t} + y_{C,t} \tag{3}$$

$$y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BC,t} + y_{CA,t} \tag{4}$$

$$y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t} \tag{5}$$

Assim, o agregado nacional pode ser representado apenas pelos agregados dos estados, através da Equação (3), ou como o agregado dos municípios (4). Já o agregado para o estado do Espírito Santo é representado por (5).

Alternativamente, podemos descrever a estrutura completa de forma matricial:

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{C,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \\ y_{BC,t} \\ y_{CA,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \\ y_{BC,t} \\ y_{CA,t} \end{bmatrix} \tag{6}$$

Por outro lado, o PIB pode ser também desagregado de forma cruzada de acordo com a atividade econômica — agricultura, indústrias extrativas, indústria de transformação, eletricidade e gás, construção etc. Essa estrutura não pode ser desagregada naturalmente de uma única forma, como é a hierarquia de estados e municípios. Não pode ser aninhada por um atributo como a própria geografia. A esse tipo de estrutura dá-se o nome de séries agrupadas.

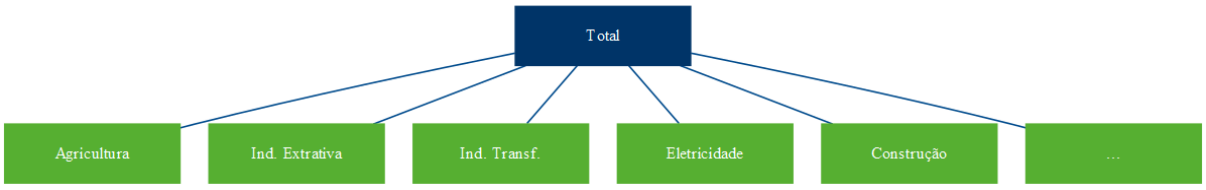

Figura 2 – Séries Agrupadas

Combinando as duas, temos a estrutura de séries hierárquicas agrupadas. Ao contrário da estrutura hierárquica, que só pode ser agregada de uma forma — como com os municípios abaixo dos estados —, a adição da estrutura agrupada pode ocorrer tanto acima (Figura 3) quanto abaixo (Figura 4) da hierárquica.

Na notação matricial, a estrutura da Figura 4 é representada como abaixo. Formalmente, o primeiro membro da igualdade é composto pelo vetor $\boldsymbol{y}_t$ n-dimensional com todas as observa-

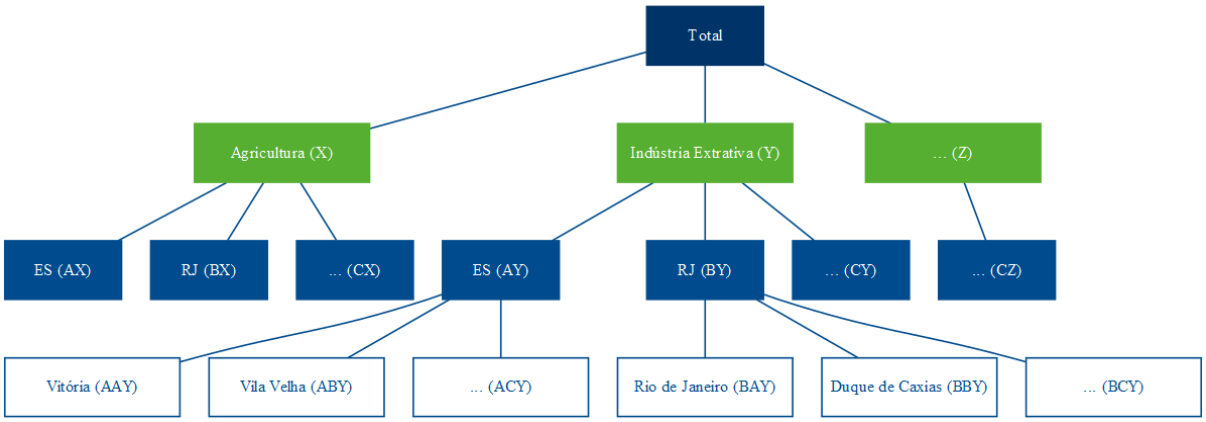

Figura 3 – Séries Hierárquicas Agrupadas (a)

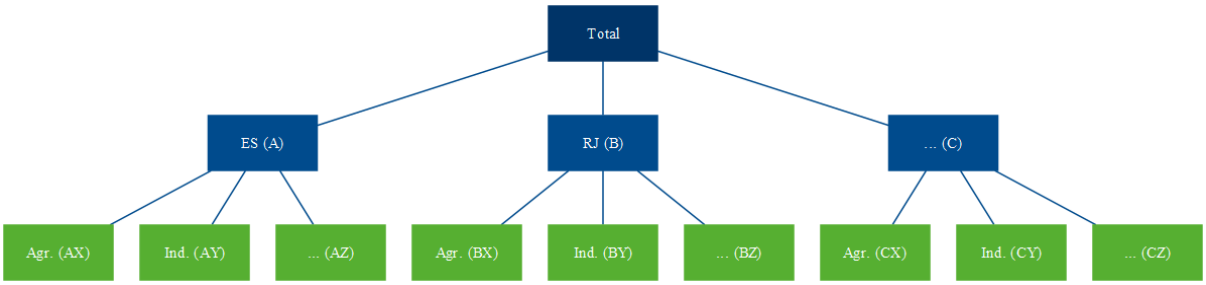

Figura 4 – Séries Hierárquicas Agrupadas (b)

ções no tempo t para todos os níveis da hierarquia. O segundo membro é composto pela matriz de soma $\boldsymbol{S}$ de dimensão $n \times m$ que define as equações para todo nível de agregação, e pela matriz $\boldsymbol{b}_t$ composta pelas séries no nível mais desagregado.

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t$$

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{C,t} \\ y_{X,t} \\ y_{Y,t} \\ y_{Z,t} \\ y_{AX,t} \\ y_{AY,t} \\ y_{AZ,t} \\ y_{BX,t} \\ y_{BY,t} \\ y_{BZ,t} \\ y_{CX,t} \\ y_{CY,t} \\ y_{CZ,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AX,t} \\ y_{AY,t} \\ y_{AZ,t} \\ y_{BX,t} \\ y_{BY,t} \\ y_{BZ,t} \\ y_{CX,t} \\ y_{CY,t} \\ y_{CZ,t} \end{bmatrix} \tag{7}$$

3.1.1 Abordagens top-down, bottom-up e middle-out

Talvez as formas mais intuitivas de se pensar em previsões para esses tipos de estrutura sejam as abordagens top-down e bottom-up. Tome a estrutura descrita na Figura 1, por exemplo. Podemos realizar a previsão para o horizonte de tempo h do agregado do PIB brasileiro, representado no topo da hierarquia por *Total* (Equação 8), e então distribuir os valores previstos proporcionalmente entre os estados e municípios.

$$\hat{\boldsymbol{y}}_{T+h|T} = E[\boldsymbol{y}_{T+h}|\Omega_T] \tag{8}$$

Essa é a abordagem top-down. Nela, a previsão para os níveis mais desagregados da hierarquia são determinadas por uma proporção p_i do nível agregado. Por exemplo, as previsões para Vitória são dadas pela Equação 9.

$$\tilde{\boldsymbol{y}}_{AA,T+h|T} = p_1 \hat{\boldsymbol{y}}_{T+h|T} \tag{9}$$

Para isso, temos de definir uma matriz com todos esses pesos, que, seguindo a formulação de Hyndman e Athanasopoulos (2021), vamos chamar de $\boldsymbol{G}$:

$$\boldsymbol{G} = \begin{bmatrix} p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{10}$$

$\boldsymbol{G}$ é uma matriz $m \times n$ que multiplica o vetor $\hat{\boldsymbol{y}}_{T+h|T}$ que chamamos de *previsões base* — as previsões individuais para todos os níveis de agregação. A equação para a abordagem *top-down* será, então:

$$\tilde{\boldsymbol{y}}_{T+h|T} = \boldsymbol{SG}\hat{\boldsymbol{y}}_{T+h|T} \tag{11}$$

Na notação matricial para a estrutura da Figura 1, temos:

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \\ \tilde{y}_{C,t} \\ \tilde{y}_{AA,t} \\ \tilde{y}_{AB,t} \\ \tilde{y}_{AC,t} \\ \tilde{y}_{BA,t} \\ \tilde{y}_{BB,t} \\ \tilde{y}_{BC,t} \\ \tilde{y}_{CA,t} \end{bmatrix} = \boldsymbol{S} \begin{bmatrix} p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{y}_{T+h|T} \\ \hat{y}_{A,T+h|T} \\ \hat{y}_{B,T+h|T} \\ \hat{y}_{C,T+h|T} \\ \hat{y}_{AA,T+h|T} \\ \hat{y}_{AB,T+h|T} \\ \hat{y}_{AC,T+h|T} \\ \hat{y}_{BA,T+h|T} \\ \hat{y}_{BB,T+h|T} \\ \hat{y}_{BC,T+h|T} \\ \hat{y}_{CA,T+h|T} \end{bmatrix} \tag{12}$$

O que nos dá uma proporção do total para cada elemento no nível mais desagregado.

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \\ \tilde{y}_{C,t} \\ \tilde{y}_{AA,t} \\ \tilde{y}_{AB,t} \\ \tilde{y}_{AC,t} \\ \tilde{y}_{BA,t} \\ \tilde{y}_{BB,t} \\ \tilde{y}_{BC,t} \\ \tilde{y}_{CA,t} \end{bmatrix} = \boldsymbol{S} \begin{bmatrix} p_1\hat{y}_{T+h|T} \\ p_2\hat{y}_{T+h|T} \\ p_3\hat{y}_{T+h|T} \\ p_4\hat{y}_{T+h|T} \\ p_5\hat{y}_{T+h|T} \\ p_6\hat{y}_{T+h|T} \\ p_7\hat{y}_{T+h|T} \end{bmatrix} \tag{13}$$

Substituindo a matriz $\boldsymbol{S}$, temos as equações que definem cada previsão da estrutura em função de proporções da previsão do agregado.

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \\ \tilde{y}_{C,t} \\ \tilde{y}_{AA,t} \\ \tilde{y}_{AB,t} \\ \tilde{y}_{AC,t} \\ \tilde{y}_{BA,t} \\ \tilde{y}_{BB,t} \\ \tilde{y}_{BC,t} \\ \tilde{y}_{CA,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_1\hat{y}_{T+h|T} \\ p_2\hat{y}_{T+h|T} \\ p_3\hat{y}_{T+h|T} \\ p_4\hat{y}_{T+h|T} \\ p_5\hat{y}_{T+h|T} \\ p_6\hat{y}_{T+h|T} \\ p_7\hat{y}_{T+h|T} \end{bmatrix} \tag{14}$$

Já a abordagem bottom-up parte do raciocínio inverso e define as previsões de cada elemento da estrutura a partir das previsões dos elementos mais desagregados. Para tanto, basta modificar a matriz $\boldsymbol{G}$.

$$\boldsymbol{G} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

O que resulta nas equações desejadas. Portanto, $\boldsymbol{G}$ define a abordagem — se *top-down* ou *bottom-up* —, e $\boldsymbol{S}$ define a maneira da qual as previsões são somadas para formar as equações de previsão para cada elemento da estrutura. Portanto, chamamos $\boldsymbol{G}$ de matriz de reconciliação.

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \\ \tilde{y}_{C,t} \\ \tilde{y}_{AA,t} \\ \tilde{y}_{AB,t} \\ \tilde{y}_{AC,t} \\ \tilde{y}_{BA,t} \\ \tilde{y}_{BB,t} \\ \tilde{y}_{BC,t} \\ \tilde{y}_{CA,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{y}_{AA,T+h|T} \\ \hat{y}_{AB,T+h|T} \\ \hat{y}_{AC,T+h|T} \\ \hat{y}_{BA,T+h|T} \\ \hat{y}_{BB,T+h|T} \\ \hat{y}_{BC,T+h|T} \\ \hat{y}_{CA,T+h|T} \end{bmatrix} \tag{16}$$

Quando m — a quantidade de elementos do nível mais desagregado — é muito grande, tornando muito custoso obter $\hat{\boldsymbol{y}}_{\boldsymbol{t}}$, e não se deseja uma abordagem estritamente *top-down*, pode-se combinar as duas formas. Ainda na estrutura hierárquica descrita na Figura 1, obter de forma criteriosa modelos Arima, por exemplo, para cada um dos municípios é muito custoso em tempo. Por outro lado, pode-se realizar a previsão para os estados e então obter de maneira *top-down* as previsões para os municípios, enquanto o nível mais agregado é obtido de maneira *bottom-up*.

$$\boldsymbol{G} = \begin{bmatrix} 0 & p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & p_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{17}$$

Esse método é chamado de *middle-out*. Nele, o total é o somatório das proporções de um nível intermédiário escolhido, ao invés de proporções do total. Isso permite uma abordagem mais econômica, em termos de custo computacional e de tempo, ao mesmo tempo em que mantém em algum grau as características individuais das hierarquias.

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \\ \tilde{y}_{C,t} \\ \tilde{y}_{AA,t} \\ \tilde{y}_{AB,t} \\ \tilde{y}_{AC,t} \\ \tilde{y}_{BA,t} \\ \tilde{y}_{BB,t} \\ \tilde{y}_{BC,t} \\ \tilde{y}_{CA,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_1\hat{y}_{A,T+h|T} \\ p_2\hat{y}_{A,T+h|T} \\ p_3\hat{y}_{A,T+h|T} \\ p_4\hat{y}_{B,T+h|T} \\ p_5\hat{y}_{B,T+h|T} \\ p_6\hat{y}_{B,T+h|T} \\ p_7\hat{y}_{C,T+h|T} \end{bmatrix} \tag{18}$$

3.1.2 Coerência e reconciliação

Seja somando as previsões do nível mais desagregado para formar os níveis superiores da hierarquia (*bottom-up*) ou distribuindo proporcionalmente as previsões do nível mais agregado (*top-down*), o vetor $\tilde{\boldsymbol{y}}_t$ representa as previsões *coerentes*. Isso significa que as previsões são totalizadas corretamente — as previsões de cada elemento agregado corresponde ao somatório das previsões dos níveis inferiores da hierarquia. Isso é garantido pela multiplicação das matrizes $\boldsymbol{SG}$.

Não fosse essa pré multiplicação, nada garantiria a coerência das previsões. Tomando a estrutura da Figura 1 como exemplo, seria um acaso improvável que as previsões do agregado para o estado do Espírito Santo sejam exatamente a soma das previsões individuais de seus municípios. Isso porque não há qualquer razão para que cada série siga o mesmo processo (e.g., arima) com coeficientes idênticos.

Os métodos de gerar previsões coerentes a partir de previsões base são chamados de métodos de *reconciliação*. Os métodos de reconciliação tradicionais apresentados, *top-down*, *bottom-up* e *middle-out*, utilizam informação limitada. No método *top-down*, utiliza-se apenas informações do nível mais agregado — por isso, apenas a primeira coluna em (Equação 10) é diferente de zero. Já na abordagem *bottom-up*, utiliza-se apenas as informações dos níveis mais desagregados, o que resulta na submatriz identidade $m \times m$ em (Equação 15), enquanto as colunas que representam os níveis mais agregados são nulas. Por fim, a abordagem *middle-out* não utiliza a mesma, mas utiliza a mesma quantidade de informação que a *top-down* (Equação 17).

Alternativamente, podemos pensar numa matriz $\boldsymbol{G}$ qualquer que utilize toda a informação disponível e tenha algumas propriedades que garantam que as previsões coerentes tenham o menor erro o possível. Esse é o problema de pesquisa trabalhado na *reconciliação ótima*.

3.2 Métodos analíticos de reconciliação ótima

[Descrever as variações do MinT: OLS, WLS e shrink]

3.3 Métodos de reconciliação ótima baseados em aprendizado de máquina

[Descrever o algorítmo, os learners — GLMnet para Lasso e Ridge, XGBoost, SVM e rede neural —, o espaço de hyperparâmetros para cada]

3.3.1 O processo de ajuste e sobreajuste

Dada uma função de ajuste f, um conjunto de pontos $D = d_1, ..., d_n$ com $d_i = (\mathbf{x}_i\ \mathbf{y}_i)'$, variáveis de decisão ou parâmetros $\mathbf{x}_i \in \mathbb{R}^m$ e imagem $\mathbf{y}_i = f(\mathbf{x}_i) \in \mathbb{R}$. Diferentemente da abordagem do modelo clássico de regressão linear em que, cumpridas certas hipóteses, há um modelo teórico de coeficientes estimados por mínimos quadrados ordinários (MQO) que é garantido pelo teorema de Gauss-Markov ser o melhor estimador linear não viesado (BLUE), em *machine learning* o objetivo é encontrar, de forma iterativa, um modelo que melhor aproxima a função f usando a informação contida em D, ou seja, queremos ajustar uma função de regressão $\hat{f}_D$ aos nossos dados D de forma que $\hat{\mathbf{y}} = \hat{f}_D(\mathbf{x}, \varepsilon)$ tenha o menor erro de aproximação ε (BISCHL et al., 2012).

Para verificar o quão bem o modelo $\hat{f}_D$ se aproxima da função real f, é necessário uma função de perda $L(\mathbf{y}, \hat{f}(\mathbf{x}))$ que, no caso de regressão, será a perda quadrática $(\mathbf{y} - \hat{f}(\mathbf{x}))^2$ ou a perda absoluta $|\mathbf{x} - \hat{f}(\mathbf{x})|$. Esses valores são agregados pela média para formar as funções de custo erro médio quadrático (MSE) e erro médio absoluto (MAE).

Calculando o custo sobre a amostra D usada para ajustar o modelo, teremos o chamado erro de resubstituição (Equação 19). Nesse caso, estaríamos usando o mesmo conjunto de dados tanto para treinar o preditor quanto para estimar o erro, o que nos levaria a uma estimativa enviesada do erro de generalização. Caso usássemos essa estimativa para seleção de modelos, esse viés favoreceria modelos mais adaptados à amostra.

$$\widehat{GE}(\hat{f}_D, D) = \sum_{(x\ y)' \in D} \frac{L(y, \hat{f}_D(x))}{|D|} \tag{19}$$

Dadas suficientes iterações, o erro de resubstituição tende a zero. Isso acontece porque conforme o preditor se adapta cada vez mais aos dados de treinamento ele irá memorizar a relação entre o conjunto de pontos D e a imagem $f(\mathbf{x}_i)$, ou seja, irá se ajustar perfeitamente ao formato da função a ser modelada. E não necessariamente um modelo perfeitamente ajustado se traduz na capacidade de predição de dados futuros (fora da amostra). De forma geral, espera-se que o preditor reduza seu viés durante o treino apenas o suficiente para que seja capaz de generalizar sua predição para fora da amostra em um nível ótimo de acurácia. A partir desse ponto, a redução no viés é penalizada com o aumento da variância, ou seja, com a redução

de sua capacidade de prever dados futuros. A esse processo se dá o nome de *overfitting* ou sobreajuste. Isso quer dizer que não podemos considerar a performance do preditor em D se desejamos estimar honestamente a performance real do modelo (BISCHL et al., 2012).

3.3.2 Reamostragem

Uma forma de se corrigir esse problema é dividindo a amostra em um conjunto para treino D^{ϕ} e outro conjunto para teste D^{*} de forma que $D^{\phi} \cup D^{\theta} = D$ e $D^{\phi} \cap D^{\theta} = \emptyset$. Assim, pode-se treinar um meta-modelo em D^{ϕ} para se obter $\hat{f}_{D^{\phi}}$ e calcular seu erro de generalização usando os dados de D^{θ}. Essa abordagem é chamada de *hold-out* e ela é de simples implementação e utilização, uma vez que as observações do conjunto teste são completamente independentes das observações com as quais o modelo foi treinado. Então, podemos estimar o erro de generalização do modelo, que consiste no cômputo do custo de $\hat{f}_{D}^{\phi}$ aplicada à amostra de teste D^{θ}.

$$\widehat{GE}(\hat{f}_{D^{\phi}}, D^{\theta}) = \sum_{(x\ y)' \in D^{\theta}} \frac{L(y, \hat{f}_{D^{\phi}}(x))}{|D^{\theta}|} \tag{20}$$

Como esse método mais simples pode não ser suficiente para detectar a variância e instabilidade de modelos mais complexos, foram desenvolvidas diferentes técnicas de reamostragem ao longo do tempo. Uma das mais populares é a validação cruzada (STONE, 1974), que consiste em gerar repetidamente i subconjuntos de treino D_i^{ϕ} e teste D_i^{θ} com o dataset disponível, ajustar um meta-modelo com cada conjunto de treino e atestar sua qualidade no conjunto de teste correspondente. A estimativa do erro de generalização então se torna:

$$\widehat{GE} = \frac{1}{k} \sum_{i=1}^{k} \widehat{GE}(\hat{f}_{D_i^{\phi}}, D_i^{\theta}) \tag{21}$$

Dividindo a amostra em k subconjuntos, utilizando $k-1$ para ajustar um meta-modelo e validando a performance no subconjunto restante — e repetindo esse procedimento para todas as possibilidades de subconjuntos —, temos a validação cruzada *k-fold*. A validação cruzada *k-fold* é uma técnica de reamostragem que permite estimar o erro de generalização de um modelo de forma mais robusta e confiável que o *hold-out* simples. Isso porque, ao contrário do *hold-out*, a validação cruzada permite que todos os dados sejam usados tanto para treino quanto para teste, o que reduz a variância da estimativa do erro de generalização.

3.3.3 Validação cruzada k-fold em séries temporais

Para dados *cross-section*, essa definição de validação cruzada *k-fold* é suficiente para qualquer caso. Entretanto, no caso de séries temporais, o analista deve tomar alguns cuidados na escolha da abordagem de validação cruzada. Isso porque, ao contrário de dados *cross-section*,

os dados de séries temporais são dependentes no tempo. Isso traz dois aparentes problemas: primeiramente, ao dividir a amostra em k subconjuntos aleatórios, o meta-modelo será treinado em períodos descontínuos e com dados futuros aos dados de teste em, ao menos, $k-1$ subconjuntos. Em segundo lugar, realizando o *split* treino-teste em y_t, com y_{t-1} na amostra treino, significa que os subconjuntos de treino e teste são dependentes (Figura 5).

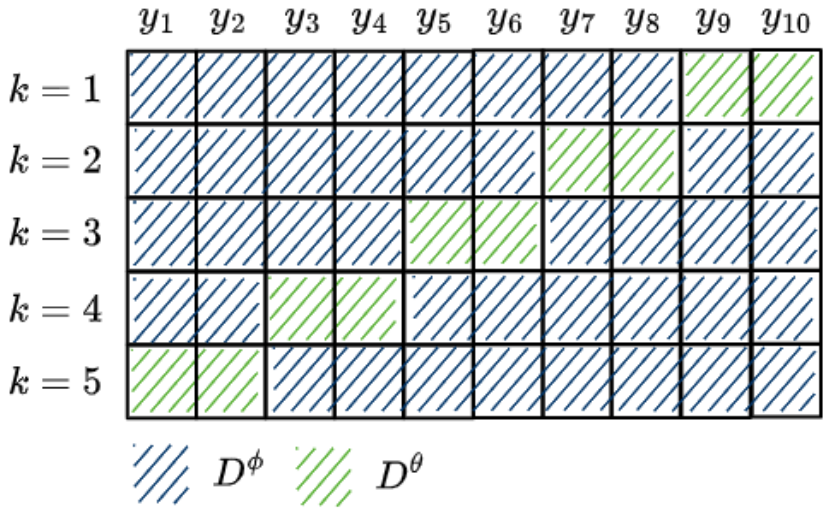

Figura 5 – Validação k-fold aleatória

Para contornar esse problema, outros métodos de validação cruzada foram desenvolvidos pensando em séries temporais. Tomando o conjunto de validação com dados exclusivamente posteriores aos dados de treino, temos o método conhecido como validação cruzada com origem móvel (Figura 6).

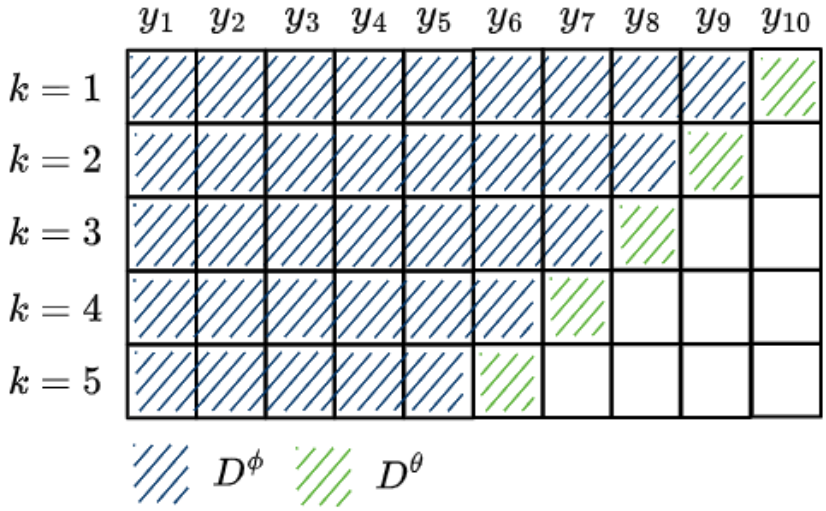

Figura 6 – Validação k-fold com origem móvel

Já quando excluímos as defasagens dependentes do subconjunto de treino (em relação ao conjunto de validação), temos a validação cruzada não dependente (Figura 7). Para ilustrar esse caso, tome um processo AR(3). Excluir as defasagens dependentes significa que, se o conjunto

de validação começa em y_t, então o conjunto de treino pode apenas ir até y_{t-4}. O problema evidente dessa abordagem é que, dependendo do tamanho da estrutura de autocorrelação da série, muitas defasagens são excluídas, podendo inviabilizar o processo de validação.

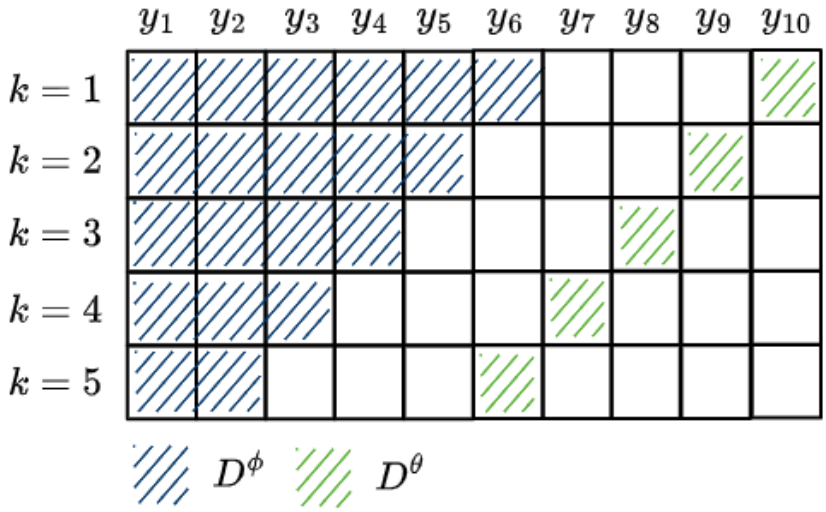

Figura 7 – Validação k-fold não-dependente

Entretanto, esses problemas são apenas aparentes. Bergmeir e Benítez (2012) comparam uma série de métodos de validação em séries temporais e concluem que, considerando séries estacionárias, os problemas teóricos relacionados à dependência não produzem impactos empíricos. Bergmeir, Hyndman e Koo (2018) vão além e abordam esse problema teórica e empiricamente, concluindo que não apenas é possível o uso de validação cruzada *k-fold* quando a série não apresenta autocorrelação serial nos resíduos, mas também é uma melhor escolha do que a validação fora da amostra[7].

4 METODOLOGIA

Neste capítulo estão contidas explicações sobre os dados e variáveis, sobre o *design* da modelagem e sobre a avaliação dos modelos.

4.1 Dados e variáveis

Os dados usados nesse trabalho são dados terciários obtidos do *datalake* público Base dos Dados (DAHIS et al., 2022). A fonte primária são os bancos comerciais e múltiplos com carteira comercial que disponibilizam mensalmente os saldos dos principais verbetes do balancete via documento 4500[8] ao Banco Central do Brasil, que os compila e publica, agrupados por

[7] *Out-of-sample evaluation* é o método padrão na literatura de séries temporais e consiste na separação da porção final da série — geralmente entre 20% e 30% — para validação.

[8] Esses documentos são relatórios eletrônicos obrigatórios demandados pelo Bacen às instituições financeiras que permitem ao regulador o conhecimento minucioso dos bancos e de seus clientes.

agência bancária e por município, no relatório ESTBAN — Estatística Bancária Mensal e por Município[9].

Além das estatísticas bancárias, foram obtidos informações de regiões, mesorregiões e microrregiões dos estados, também a partir *datalake* Base dos Dados, com o objetivo de enriquecer a estrutura hierárquica dos dados do ESTBAN, limitada aos municípios.

Uma vez que o escopo deste trabalho se encerra ao Espírito Santo e ao Banestes, foram aplicados os filtros para UF e na raiz do CNPJ. Ademais, foram mantidos apenas os verbetes relacionados a crédito e mantidas apenas as agências em atividade durante todo o período. Quanto ao período, há dados disponíveis desde 1988. Entretanto, escolhi manter apenas os dados a partir de 2010 pois, se tratando de uma hierarquia larga, o custo computacional deve ser levado em conta.

Por fim, as variáveis mantidas no *dataset* foram:

1. ref: data de referência do relatório ESTBAN
2. nome_mesorregiao: nome da mesorregião do ES:
    - Central Espírito-Santense
    - Litoral Norte Espírito-Santense
    - Noroeste Espírito-Santense
    - Sul Espírito-Santense
3. nome_microrregião: nome da microrregião do ES:
    - Afonso Cláudio
    - Guarapari
    - Santa Teresa
    - Vitória
    - Linhares
    - Montanha
    - São Mateus
    - Barra de São Francisco
    - Colatina
    - Nova Venécia
    - Alegre
    - Cachoeiro de Itapemirim
    - Itapemirim
4. verbete:
    - empréstimos e títulos descontados

[9] https://www4.bcb.gov.br/fis/cosif/estban.asp?frame=1

- financiamentos
- financiamentos imobiliários
- financiamentos rurais

5. nome: nome do município
6. cnpj_agencia
7. valor: saldo do verbete no município

Os dados foram organizados de forma hierárquica, do mais agregado para o mais desagregado, por estado, mesorregião, microrregião, município e agência bancária; e de forma agrupada, por verbete.

O detalhamento da construção do *dataset* se encontra no Anexo A.

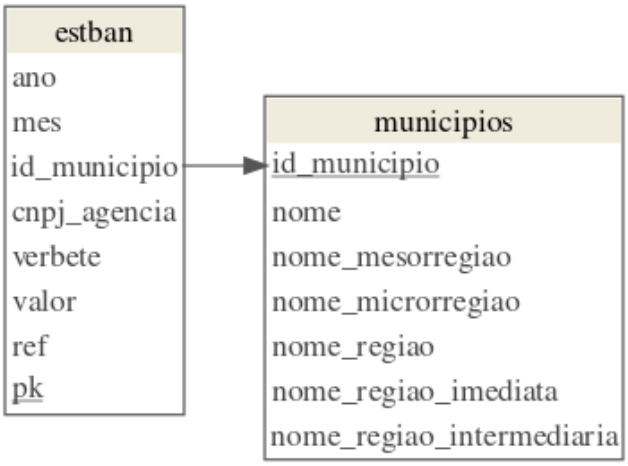

Figura 8 – Modelo de dados

4.2 Análise exploratória dos dados

Após limpeza, o *dataset* adquiriu a estrutura apresentada na Tabela 1. As microrregiões que compõem cada mesorregião são apresentadas na Tabela 2 e os municípios que compõem cada microrregião são apresentados na Tabela 3.

Tabela 1 – Estrutura do dataset

ref	nome_mesorregiao	nome_microrregiao	nome	cnpj_agencia	verbete	valor
2023-01-01	Central Espírito-santense	Vitória	Vitória	28127603014802	financiamentos imobiliários	1449593453
2023-01-01	Central Espírito-santense	Vitória	Vitória	28127603014802	financiamentos rurais	651690
2023-01-01	Central Espírito-santense	Vitória	Vitória	28127603016767	empréstimos e títulos descontados	29341002
2023-01-01	Central Espírito-santense	Vitória	Vitória	28127603016767	financiamentos	686391
2023-01-01	Central Espírito-santense	Vitória	Vitória	28127603016767	financiamentos imobiliários	0
2023-01-01	Central Espírito-santense	Vitória	Vitória	28127603016767	financiamentos rurais	1485660

O tamanho de uma estrutura hierárquica, em termos de observações, é determinada por seu nível mais desagregado. Assim, sendo 155 meses e 96 agências, a estrutura hierárquica conta com $155 \times 96 = 14880$ observações. Sendo também uma estrutura agrupada por 4 verbetes, a

Tabela 2 – Microrregiões do ES incluídas nos dados

nome_mesorregiao	nome_microrregiao
Central Espírito-santense	Afonso Cláudio, Guarapari , Santa Teresa , Vitória
Litoral Norte Espírito-santense	Linhares , Montanha , São Mateus
Noroeste Espírito-santense	Barra de São Francisco, Colatina , Nova Venécia
Sul Espírito-santense	Alegre , Cachoeiro de Itapemirim, Itapemirim

Tabela 3 – Municípios por microrregião do ES incluídos nos dados

nome_microrregiao	nome
Afonso Cláudio	Afonso Cláudio , Brejetuba , Conceição do Castelo , Domingos Martins , Laranja da Terra , Marechal Floriano , Venda Nova do Imigrante
Alegre	Alegre , Dores do Rio Preto, Guaçuí , Ibatiba , Ibitirama , Irupi , Iúna , Muniz Freire
Barra de São Francisco	Água Doce do Norte , Barra de São Francisco, Ecoporanga , Mantenópolis
Cachoeiro de Itapemirim	Apiacá , Atilio Vivacqua , Bom Jesus do Norte , Cachoeiro de Itapemirim, Castelo , Jerônimo Monteiro , Mimoso do Sul , Muqui , São José do Calçado , Vargem Alta
Colatina	Alto Rio Novo, Baixo Guandu , Colatina , Pancas
Guarapari	Alfredo Chaves , Anchieta , Guarapari , Iconha , Piúma , Rio Novo do Sul
Itapemirim	Itapemirim , Marataízes , Presidente Kennedy
Linhares	Aracruz , Fundão , Ibiraçu , João Neiva , Linhares , Rio Bananal, Sooretama
Montanha	Mucurici , Pinheiros , Ponto Belo
Nova Venécia	Águia Branca , Boa Esperança, Nova Venécia , Vila Valério
Santa Teresa	Itaguaçu , Itarana , Santa Leopoldina , Santa Maria de Jetibá, Santa Teresa , São Roque do Canaã
São Mateus	Jaguaré , Pedro Canário, São Mateus
Vitória	Cariacica , Serra , Viana , Vila Velha, Vitória

quantidade de observações é multiplicada pela quantidade de níveis transversais, totalizando 59520 observações.

Tabela 4 – Contagem de únicos no dataset ESTBAN

	unique_n
ref	155
nome_mesorregiao	4
nome_microrregiao	13
nome	70
cnpj_agencia	96
verbete	4

A série temporal do agregado de crédito no Banestes no Espírito Santo é apresentada na Figura 9. Podemos observar tendência de crescimento a partir de 2020, indicando que a série seja não estacionária e que talvez seja interessante adicionar um regressor externo para o período da pandemia, quando foram implementados programas de crédito emergenciais no país.

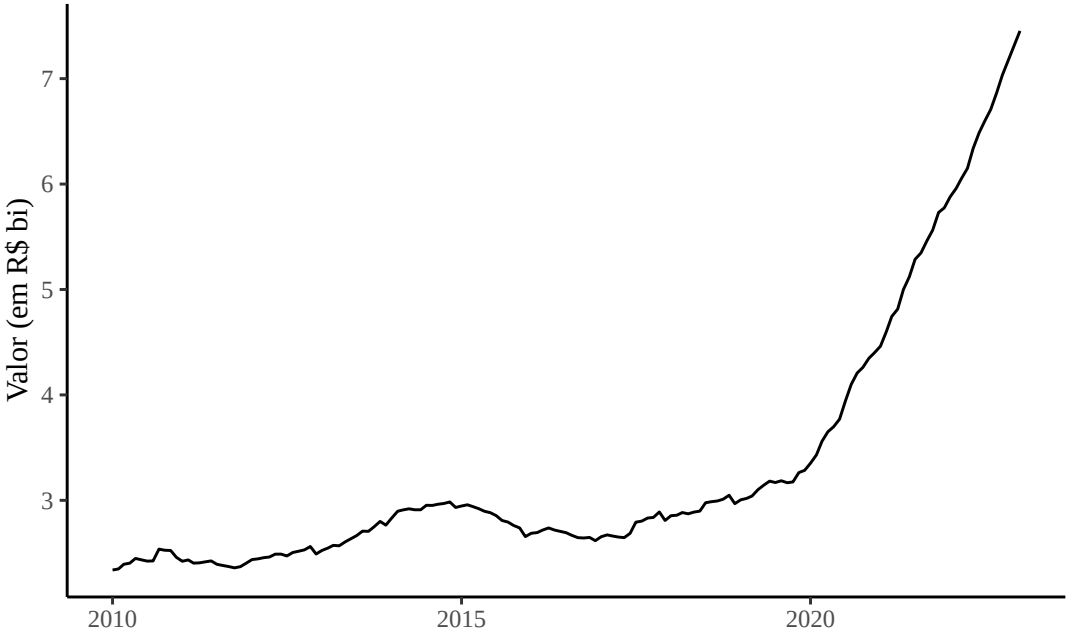

Figura 9 – Série temporal do agregado de crédito do Banestes no ES

A concentração na mesorregião Central Espírito-santense, Figura 10, indica que os elementos das demais regiões

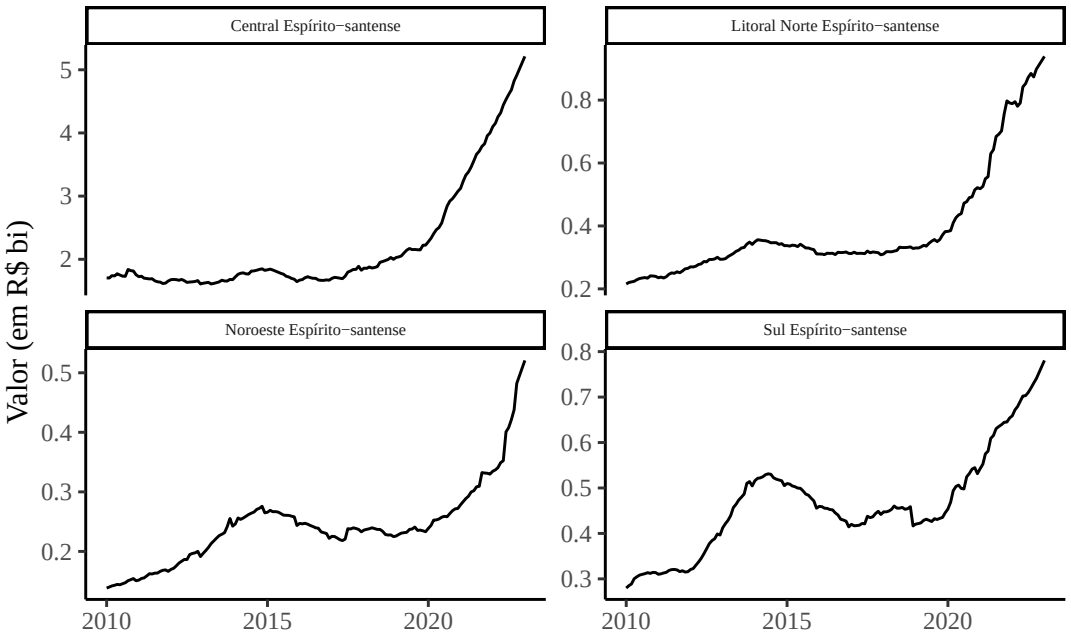

Figura 10 – Série temporal do agregado de crédito do Banestes por mesorregião do ES

Analisando o agregado por verbetes, a Figura 12 indica que o crescimento é liderado pelas alíneas empréstimos e títulos descontados e financiamentos imobiliários. Especificamente quanto ao segundo, a competitividade da taxa de juros é um fator que pode ter contribuído para o crescimento, o que também pode ser analisado com o auxílio de um regressor externo.

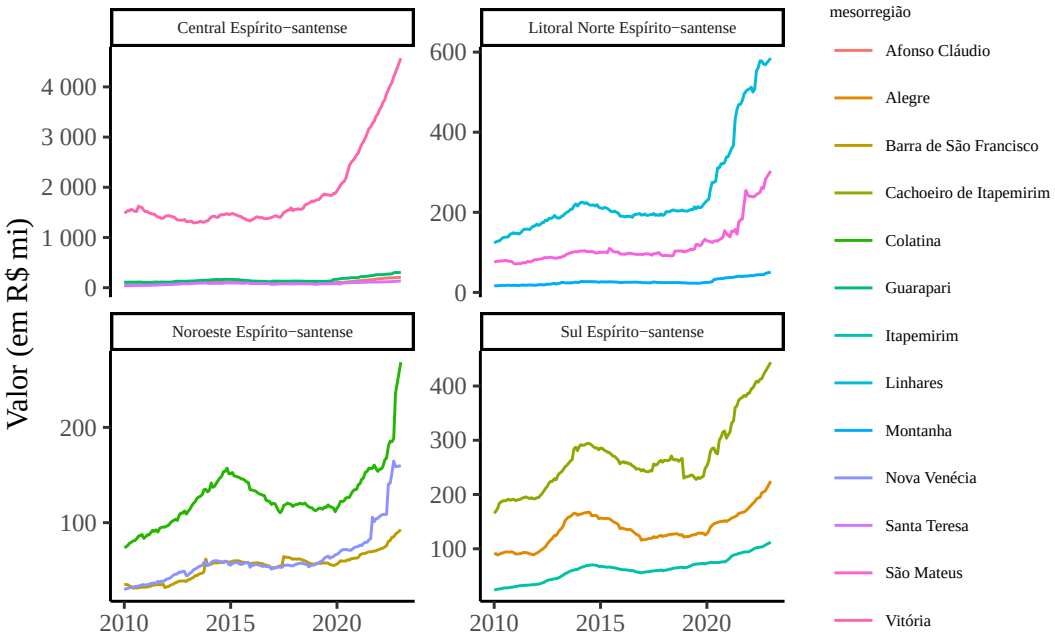

Figura 11 – Série temporal do agregado de crédito do Banestes por microrregião do ES

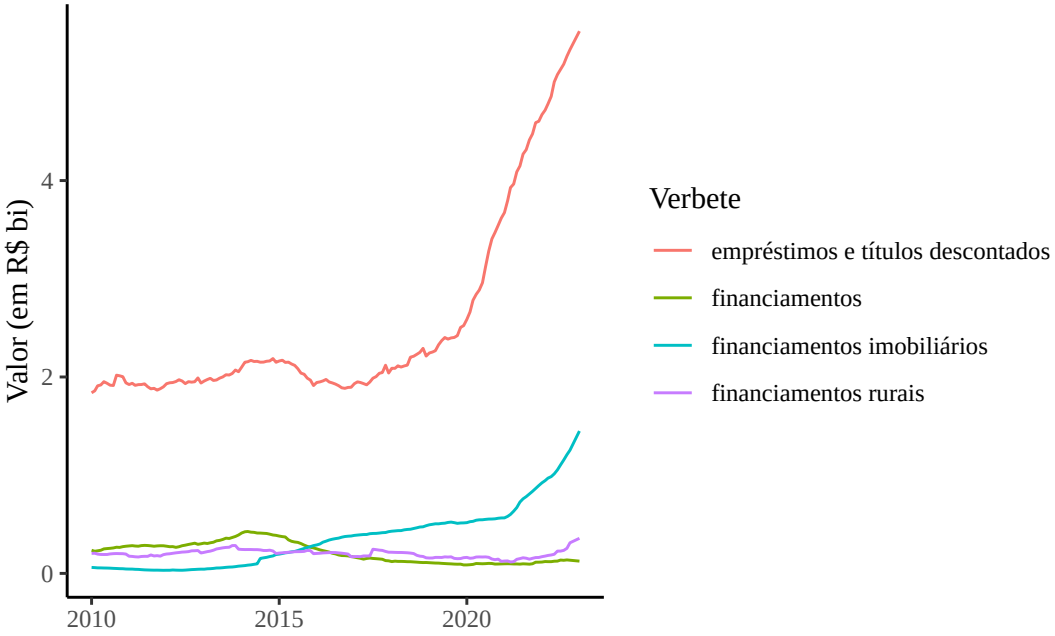

Figura 12 – Verbetes no agregado do ES

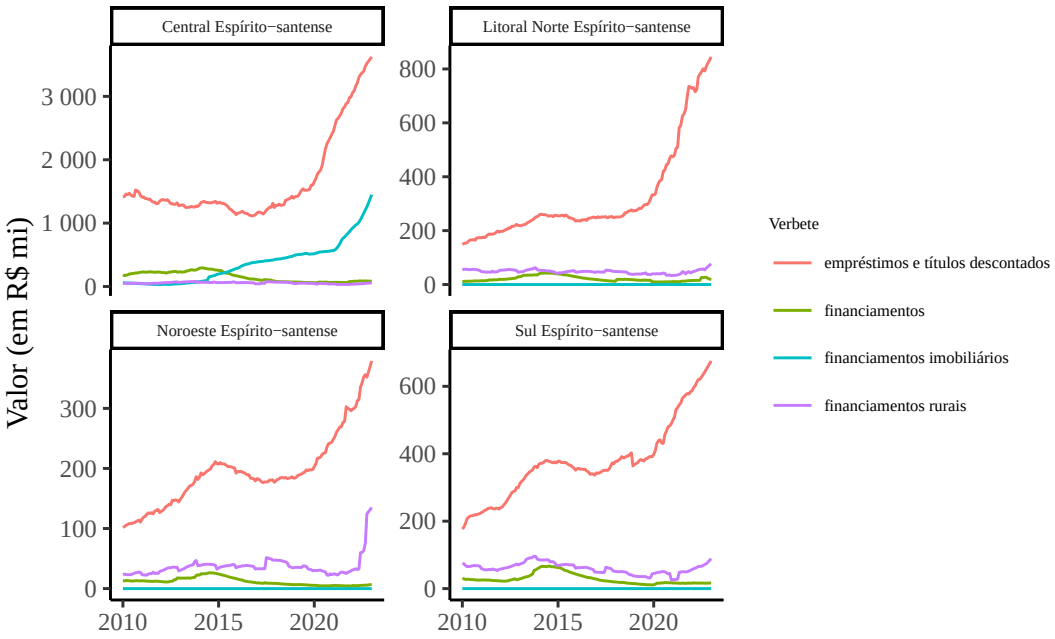

Figura 13 – Verbete por mesorregião do ES

5 RESULTADOS

[Escrever *outline* do capítulo]

REFERÊNCIAS

ATHANASOPOULOS, George; AHMED, Roman A.; HYNDMAN, Rob J. Hierarchical forecasts for Australian domestic tourism. en. *International Journal of Forecasting*, v. 25, n. 1, p. 146–166, jan. 2009. ISSN 0169-2070. DOI: 10.1016/j.ijforecast.2008.07.004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207008000691>. Acesso em: 11 jan. 2023. Citado na p. 16.

BADER, Fani Lea Cymrot; KOYAMA, SÉRGIO MIKIO; TSUCHIDA, MARCOS HIROYUKI. Modelo FAVAR Canônico para Previsão do Mercado de Crédito. pt. *Banco Central do Brasil*, v. 369, p. 38, nov. 2014. ISSN 1519-1028. Citado na p. 14.

BECCALLI, Elena et al. Earnings management, forecast guidance and the banking crisis. *The European Journal of Finance*, v. 21, n. 3, p. 242–268, fev. 2015. Publisher: Routledge _eprint: https://doi.org/10.1080/1351847X.2013.809548. ISSN 1351-847X. DOI: 10.1080/1351847X.2013.809548. Disponível em: <https://doi.org/10.1080/1351847X.2013.809548>. Acesso em: 7 mai. 2023. Citado na p. 12.

BERGMEIR, Christoph; BENÍTEZ, José M. On the use of cross-validation for time series predictor evaluation. en. *Information Sciences*, v. 191, p. 192–213, mai. 2012. ISSN 0020-0255. DOI: 10.1016/j.ins.2011.12.028. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025511006773>. Acesso em: 30 mai. 2023. Citado na p. 29.

BERGMEIR, Christoph; HYNDMAN, Rob J.; KOO, Bonsoo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. en. *Computational Statistics & Data Analysis*, v. 120, p. 70–83, abr. 2018. ISSN 01679473. DOI: 10.1016/j.csda.2017.11.003. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0167947317302384>. Acesso em: 28 mai. 2023. Citado na p. 29.

BISCHL, Bernd et al. Resampling Methods for Meta-Model Validation with Recommendations for Evolutionary Computation. *Evolutionary computation*, v. 20, p. 249–75, fev. 2012. DOI: 10.1162/EVCO_a_00069. Citado nas pp. 26, 27.

BRASIL. *Lei nº 7.492, de 16 de junho de 1986*. Brasília, DF: Presidência da República, jun. 1986. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l7492.htm>. Citado na p. 11.

COELHO, Arthur Nascimento Bernardes et al. A responsabilidade da auditoria externa na fraude contábil do banco panamericano. pt. *RAGC*, v. 3, n. 7, set. 2015. Number: 7. ISSN 2317-0484. Disponível em: <https://revistas.fucamp.edu.br/index.php/ragc/article/view/604>. Acesso em: 31 mai. 2023. Citado na p. 12.

ÇOLAK, Mehmet Selman et al. *TCMB - Monitoring and Forecasting Cyclical Dynamics in Bank Credits: Evidence from Turkish Banking Sector*. en. [S.l.: s.n.], 2019. Disponível em: <https://www.tcmb.gov.tr/wps/wcm/connect/EN/TCMB+EN/Main+Menu/Publications/Research/Working+Paperss/2019/19-29>. Acesso em: 6 mar. 2023. Citado na p. 14.

CONSELHO MONETÁRIO NACIONAL. *Resolução nº 2.099, de 17 de agosto de 1994*. Brasília, DF: Banco Central do Brasil, ago. 1994. Disponível em: <https://www.bcb.gov.br/pre/normativos/res/1994/pdf/res_2099_v1_O.pdf>. Citado na p. 12.

DAHIS, Ricardo et al. *Data Basis (Base Dos Dados): Universalizing Access to High-Quality Data*. en. Rochester, NY: [s.n.], jul. 2022. DOI: 10.2139/ssrn.4157813. Disponível em: <https://papers.ssrn.com/abstract=4157813>. Acesso em: 30 mai. 2023. Citado na p. 29.

GORODETSKAYA, Olga; GOBAREVA, Yana; KOROTEEV, Mikhail. A Machine Learning Pipeline for Forecasting Time Series in the Banking Sector. en. *Economies*, v. 9, n. 4, p. 205, dez. 2021. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 2227-7099. DOI: 10.3390/economies9040205. Disponível em: <https://www.mdpi.com/2227-7099/9/4/205>. Acesso em: 27 fev. 2023. Citado na p. 14.

GROSS, Charles W.; SOHL, Jeffrey E. Disaggregation methods to expedite product line forecasting. en. *Journal of Forecasting*, v. 9, n. 3, p. 233–254, 1990. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.3980090304. ISSN 1099-131X. DOI: 10.1002/for.3980090304. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/for.3980090304>. Acesso em: 26 jan. 2023. Citado nas pp. 13, 15.

HYNDMAN, R.J.; ATHANASOPOULOS, G. *Forecasting: principles and practice*. 3. ed. Melbourne, Austrália: OTexts, 2021. Disponível em: <https://otexts.com/fpp3/>. Citado nas pp. 13, 15, 18, 21.

HYNDMAN, Rob J.; AHMED, Roman A. et al. Optimal combination forecasts for hierarchical time series. en. *Computational Statistics & Data Analysis*, v. 55, n. 9, p. 2579–2589, set. 2011. ISSN 0167-9473. DOI: 10.1016/j.csda.2011.03.006. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947311000971>. Acesso em: 11 jan. 2023. Citado nas pp. 13, 16.

HYNDMAN, Rob J.; LEE, Alan J.; WANG, Earo. Fast computation of reconciled forecasts for hierarchical and grouped time series. en. *Computational Statistics & Data Analysis*, v. 97, p. 16–32, mai. 2016. ISSN 0167-9473. DOI: 10.1016/j.csda.2015.11.007. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016794731500290X>. Acesso em: 11 jan. 2023. Citado na p. 17.

LI, Zhaoxuan et al. A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting. en. *Energies*, v. 9, n. 1, p. 55, jan. 2016. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 1996-1073. DOI: 10.3390/

en9010055. Disponível em: <https://www.mdpi.com/1996-1073/9/1/55>. Acesso em: 8 abr. 2023. Citado na p. 16.

MOURA, Denia de. *Análise dos fatores de convencimento do juízo brasileiro quanto à ocorrência de fraude contábil: um estudo de caso Múltiplo da Gallus, da Encol e do Banco Santos*. 2007. Dissertação de Mestrado – Fundação Getúlio Vargas, Rio de Janeiro. Accepted: 2009-11-18T19:01:34Z. Disponível em: <http://bibliotecadigital.fgv.br:80/dspace/handle/10438/4038>. Acesso em: 31 mai. 2023. Citado na p. 12.

PANAGIOTELIS, Anastasios et al. Forecast reconciliation: A geometric view with new insights on bias correction. en. *International Journal of Forecasting*, v. 37, n. 1, p. 343–359, jan. 2021. ISSN 0169-2070. DOI: 10.1016/j.ijforecast.2020.06.004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207020300911>. Acesso em: 15 jan. 2023. Citado na p. 17.

PRAYOGA, I.G.S.A.; SUHARTONO, Suhartono; RAHAYU, S.P. Top-down forecasting for high dimensional currency circulation data of Bank Indonesia. *International Journal of Advances in Soft Computing and its Applications*, v. 9, p. 62–74, jan. 2017. Citado na p. 15.

SEZER, Omer Berat; GUDELEK, Mehmet Ugur; OZBAYOGLU, Ahmet Murat. *Financial Time Series Forecasting with Deep Learning : A Systematic Literature Review: 2005-2019*. [S.l.]: arXiv, nov. 2019. arXiv:1911.13288 [cs, q-fin, stat]. Disponível em: <http://arxiv.org/abs/1911.13288>. Acesso em: 7 mar. 2023. Citado na p. 14.

SPILIOTIS, Evangelos et al. Hierarchical forecast reconciliation with machine learning. en. *Applied Soft Computing*, v. 112, p. 107756, nov. 2021. ISSN 1568-4946. DOI: 10.1016/j.asoc.2021.107756. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1568494621006773>. Acesso em: 11 jan. 2023. Citado nas pp. 17, 18.

STONE, M. Cross-validation and multinomial prediction. *Biometrika*, v. 61, n. 3, p. 509–515, dez. 1974. ISSN 0006-3444. DOI: 10.1093/biomet/61.3.509. Disponível em: <https://doi.org/10.1093/biomet/61.3.509>. Acesso em: 28 mai. 2023. Citado na p. 27.

WALLANDER, Jan. Budgeting — an unnecessary evil. en. *Scandinavian Journal of Management*, v. 15, n. 4, p. 405–421, dez. 1999. ISSN 0956-5221. DOI: 10.1016/S0956-5221(98)00032-3. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0956522198000323>. Acesso em: 8 mai. 2023. Citado na p. 11.

WICKRAMASURIYA, Shanika L.; ATHANASOPOULOS, George; HYNDMAN, Rob J. Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association*, v. 114, n. 526, p. 804–819, abr. 2019. Publisher: Taylor & Francis. ISSN 0162-1459. DOI: 10.1080/01621459.2018.1448825. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1448825>. Acesso em: 11 jan. 2023. Citado nas pp. 13, 17.

Anexos

ANEXO A – CÓDIGO PARA CONSTRUÇÃO DA BASE DE DADOS

```
# pacotes
library(magrittr, include.only = "%>%")

# municípios x regiões imediatas
municipios = basedosdados::read_sql("
SELECT
  id_municipio
  , nome
  , nome_mesorregiao
  , nome_microrregiao
  , nome_regiao
  , nome_regiao_imediata
  , nome_regiao_intermediaria
FROM `basedosdados.br_bd_diretorios_brasil.municipio`
WHERE sigla_uf = 'ES'
")

# estban
estban = basedosdados::read_sql("
SELECT
  CAST(ano AS STRING) AS ano
  , CAST(mes AS STRING) AS mes
  , id_municipio
  , cnpj_agencia
  , CASE
        WHEN id_verbete = '160' THEN 'operações de crédito'
        WHEN id_verbete = '161' THEN 'empréstimos e títulos descontados'
        WHEN id_verbete = '162' THEN 'financiamentos'
        WHEN id_verbete = '163' THEN 'financiamentos rurais'
        WHEN id_verbete = '169' THEN 'financiamentos imobiliários'
        WHEN id_verbete = '172' THEN 'outros créditos'
        WHEN id_verbete = '174' THEN 'provisão para operações de crédito'
        ELSE 'outros'
    END AS verbete
  , valor
```

```
FROM `basedosdados.br_bcb_estban.agencia`

WHERE
  -- CNPJ do Banestes
  cnpj_basico = '28127603'
  -- filtrando verbetes de interesse
  AND id_verbete IN ('161', '162', '163', '169')
")

# formatando datas
estban = within(estban, {
  mes = formatC(as.numeric(mes), format = "d", width = 2, flag = "0")
  ref = as.Date(paste(ano, mes, "01", sep = "-"))
})

# identificando agências em atividade
agencias_fim = subset(estban, ref == max(ref), select = cnpj_agencia) |>
  (\(x) unique(x$cnpj_agencia))()

# filtrando apenas agências em atividade
estban = subset(
  estban,
  cnpj_agencia %in% agencias_fim
)

# mesclando com tabela municípios
estban = merge(estban, municipios, by = "id_municipio")
```