UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM DE SISTEMAS
COMPLEXOS

LUDMILA DEUTE RIBEIRO

Leis de Escala nos Gastos com Saneamento Básico: dados do SIOP e DOU

São Paulo

2019

LUDMILA DEUTE RIBEIRO

Leis de Escala nos Gastos com Saneamento Básico: dados do SIOP e DOU

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre pelo Programa de Pós-graduação em Modelagem de Sistemas Complexos.

Área de concentração: Sistemas Complexos aplicados a Ciências Sociais, Biológicas, Ambientais e de Saúde.

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 14/03/2019. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 14 de outubro de 2011.

Orientador: Prof. Dr. Camilo Rodrigues Neto

São Paulo

2019

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB 8 - 4936

Ribeiro, Ludmila Deute

Leis de Escala nos gastos com saneamento básico: dados do SIOP e DOU / Ludmila Deute Ribeiro ; orientador Camilo Rodrigues Neto. – 2019.

148 f. + 1 Pen-drive : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Modelagem de Sistemas Complexos, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo. Versão corrigida.

1. Saneamento básico - Brasil. 2. Investimentos públicos - Brasil. 3. Cidades - Brasil. 4. Redes complexas. 5. Diário Oficial da União - Análise. 6. Sistema Integrado de planejamento e Orçamento do Governo Federal - Análise. 7. Processamento de linguagem natural - Aplicações. I. Rodrigues Neto, Camilo, orient. II. Título.

CDD 22.ed.– 628.0981

Dissertação de autoria de Ludmila Deute Ribeiro, sob o título “**Leis de Escala nos Gastos com Saneamento Básico: dados do SIOP e DOU**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Modelagem de Sistemas Complexos, na área de concentração Sistemas Complexos aplicados a Ciências Sociais, Biológicas, Ambientais e de Saúde, aprovada em 14 de março de 2019 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Camilo Rodrigues Neto

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo
Presidente

Prof. Dr. José de Jesus Pérez Alcazár

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

Prof. Dr. Fernando Fagundes Ferreira

Instituição: Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

Prof. Dr. Amaury de Souza Amaral

Pontifícia Universidade Católica de São Paulo

*A todos os que se dedicam à pesquisa científica no Brasil, a despeito dos muitos
obstáculos...*

Agradecimentos

Essa dissertação de mestrado não teria sido concluída sem o apoio de várias pessoas e instituições.

Em primeiro lugar, agradeço ao meu orientador, Professor Dr. Camilo Rodrigues Neto, pela confiança depositada desde o nosso primeiro contato. E pela disposição de corrigir os meus erros, sempre que necessário.

Em segundo lugar, agradeço aos colegas e professores do Mestrado em Modelagem de Sistemas Complexos da EACH/USP, que me receberam de braços abertos e que me apoiaram durante todo o curso.

Agradeço também ao meu supervisor de estágio do Programa de Aperfeiçoamento de Ensino da USP, Professor Dr. Fernando de Souza Coelho, pela oportunidade de participar de atividades acadêmicas referentes ao curso de Gestão de Políticas Públicas.

Finalmente, agradeço à minha família e aos amigos pelo apoio incondicional, e peço desculpas pelas muitas horas subtraídas ao nosso convívio.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“O estudo dos sistemas complexos será a ciência do século XXI...”
(Stephen Hawking)

Resumo

RIBEIRO, Ludmila Deute. **Leis de Escala no Saneamento Básico**: dados do SIOP e DOU. 2019. 148 p. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2019.

A partir do final do século 20, o governo federal criou vários programas visando a ampliação de acesso ao saneamento básico. Embora esses programas tenham trazido o abastecimento de água potável e a coleta de resíduos sólidos para a maioria dos municípios brasileiros, o esgotamento sanitário ainda está espacialmente concentrado na região Sudeste e nas áreas mais urbanizadas. Para explicar esse padrão espacialmente concentrado, é frequentemente assumido que o tamanho das cidades realmente importa para o saneamento básico, especialmente para o esgotamento sanitário. De fato, à medida que as cidades crescem em tamanho, devemos esperar economias de escala no volume de infraestrutura de saneamento. Economias de escala na infra-estrutura implicam uma redução nos custos de saneamento básico, de forma proporcional ao tamanho da cidade, levando também a uma (esperada) relação de lei de escala (ou de potência) entre os gastos com saneamento básico e o tamanho da cidade. Usando a população, $N(t)$, como medida do tamanho da cidade no momento t , a lei de escala para infraestrutura assume o formato $Y(t) = Y_0 N(t)^\beta$ onde $\beta \approx 0.8 < 1$, Y denota o volume de infraestrutura e Y_0 é uma constante. Diversas propriedades das cidades, desde a produção de patentes e renda até a extensão da rede elétrica, são funções de lei de potência do tamanho da população com expoentes de escalamento, β , que se enquadram em classes distintas. As quantidades que refletem a criação de riqueza e a inovação têm $\beta \approx 1.2 > 1$ (retornos crescentes), enquanto aquelas responsáveis pela infraestrutura exibem $\beta \approx 0.8 < 1$ (economias de escala). Verificamos essa relação com base em dados extraídos do Sistema Integrado de Planejamento e Orçamento (SIOP), que abrangem transferências com recursos não onerosos, previstos na Lei Orçamentária Anual (LOA), na modalidade saneamento básico. No conjunto, os valores estimados de β mostram redução das transferências da União Federal para saneamento básico, de forma proporcional ao tamanho dos municípios beneficiários. Para a dotação inicial, valores previstos na LOA, β estimado é de aproximadamente: 0.63 para municípios com população superior a dois mil habitantes; 0.92 para municípios acima de vinte mil habitantes; e de 1.18 para municípios com mais de cinquenta mil habitantes. A segunda fonte de dados identificada foi o Diário Oficial da União (DOU), periódico do governo federal para publicação de atos oficiais. Os dados fornecidos pelo DOU referem-se aos recursos não onerosos e também aos empréstimos com recursos do Fundo de Garantia por Tempo de Serviço (FGTS). Para extração dos dados textuais foram utilizadas técnicas de Processamento de Linguagem Natural (PLN). Essas técnicas funcionam melhor quando os algoritmos são alimentados com anotações - metadados que fornecem informações adicionais sobre o texto. Por isso geramos uma base de dados, a partir de textos anotados do DOU, para treinar uma rede LSTM bidirecional aplicada à etiquetagem morfossintática e ao reconhecimento de entidades nomeadas. Os resultados preliminares obtidos desta forma estão relatados no texto.

Palavras-chave: Cidades. Saneamento Básico. Leis de Escala. Processamento de Linguagem Natural. Etiquetagem Morfossintática. Reconhecimento de Entidades Nomeadas.

Abstract

RIBEIRO, Ludmila Deute. **Scaling Patterns in Basic Sanitation Expenditure:** data from SIOP and DOU. 2019. 148 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2019.

Starting in the late 20th century, the Brazilian federal government created several programs to increase the access to water and sanitation. However, although these programs made improvements in water access, sanitation was generally overlooked. While water supply, and waste collection are available in the majority of the Brazilian municipalities, the sewage system is still spatially concentrated in the Southeast region and in the most urbanized areas. In order to explain this spatially concentrated pattern it is frequently assumed that the size of cities does really matter for sanitation services provision, specially for sewage collection. As a matter of fact, as cities grow in size, one should expect economies of scale in sanitation infrastructure volume. Economies of scale in sanitation infrastructure means a decrease in basic sanitation costs, proportional to the city size, leading also to a (expected) power law relationship between the expenditure on sanitation and city size. Using population, $N(t)$, as the measure of city size at time t , power law scaling for infrastructure takes the form $Y(t) = Y_0 N(t)^\beta$ where $\beta \approx 0.8 < 1$, Y denotes infrastructure volume and Y_0 is a constant. Many diverse properties of cities from patent production and personal income to electrical cable length are shown to be power law functions of population size with scaling exponents, β , that fall into distinct universality classes. Quantities reflecting wealth creation and innovation have $\beta \approx 1.2 > 1$ (increasing returns), whereas those accounting for infrastructure display $\beta \approx 0.8 < 1$ (economies of scale). We verified this relationship using data from federal government databases, called Integrated Planning and Budgeting System, known as SIOP. SIOP data refers only to grants, funds given to municipalities by the federal government to run programs within defined guidelines. Preliminary results from SIOP show decrease in Federal Grants to Brazilian Municipalities, proportional to the city size. For the initial budget allocation, β was found to be roughly 0.63 for municipalities above twenty thousand inhabitants; to be roughly 0.92 for municipalities above twenty thousand inhabitants; and to be roughly 1.18 for municipalities above fifty thousand inhabitants. The second data source is DOU, government journal for publishing official acts. DOU data should give us information not only about grants, but also about FGTS funds for basic sanitation loans. In order to extract data from DOU we have applied Natural Language Processing (NLP) tools. These techniques often work better when the algorithms are provided with annotations— metadata that provides additional information about the text. In particular, we fed a database with annotations into a bidirectional LSTM model applied to POS Tagging and Named-entity Recognition. Preliminary results are reported in the paper.

Keywords: Cities. Basic Sanitation. Power Law Scaling. Natural Language Processing. POS Tagging. Named-entity Recognition.

Lista de figuras

Figura 1 – Panorama dos planos municipais de saneamento básico - 2017	23
Figura 2 – Evolução dos recursos da LOA e percentual de emendas parlamentares - 2004 a 2016	28
Figura 3 – Representação espacial do índice de atendimento urbano por redes de abastecimento de água, segundo faixas percentuais, por município - 2017	36
Figura 4 – Representação espacial do índice de atendimento urbano por redes coletoras de esgotos, segundo faixas percentuais, por município - 2017 .	37
Figura 5 – Taxa de serviços de coleta de resíduos sólidos em relação à população urbana, segundo faixas populacionais - 2017	38
Figura 6 – Escalas de urbanização, segundo regiões de articulação ampliada	39
Figura 7 – Brasil: participação dos municípios no população total e no PIB - 2010	40
Figura 8 – Laço de retroalimentação o mais simples possível com uma taxa e um nível	44
Figura 9 – Estrutura dos laços de retroalimentação negativos	44
Figura 10 – Escalamento sublinear e superlinear em cidades	47
Figura 11 – Sistemas de localidades centrais de Christaller	49
Figura 12 – Gráfico tamanho-frequência dos <i>clusters</i> para a Holanda - 1994	51
Figura 13 – Modelo de Barabási-Albert: grafo	58
Figura 14 – Modelo de Barabási-Albert: histograma	58
Figura 15 – Padrões de conectividade de rede	61
Figura 16 – Distribuição dos nós da rede por classificação	61
Figura 17 – A hierarquia implícita do varejo no centro de Londres	62
Figura 18 – Extrato de convênio, 01/08/2018, município de Parnaíba- PI	76
Figura 19 – Leitura de jornais, 22/08/2018, Seção 3, MCidades	77
Figura 20 – Famílias de sistemas de apoio da Administração Pública Federal	78
Figura 21 – Arquitetura de extração de informações	87
Figura 22 – Cadeias de Markov para tempo e palavras	94
Figura 23 – Modelo oculto de Markov para consumo de sorvetes e tempo	95
Figura 24 – Perceptron multicamadas com 3 camadas	100
Figura 25 – Gradiente descendente	100

Figura 26 – Algoritmo de retropropagação em uma rede neural de 3 camadas . . .	102
Figura 27 – Adição de conexões recorrentes a uma rede <i>feedforward</i> simples . . .	104
Figura 28 – Unidades SRN e LSTM	105
Figura 29 – Segmentação e etiquetagem nos níveis dos <i>tokens</i> e dos <i>chunks</i>	106
Figura 30 – Modelo concatenado CRF-BLSTM com camada <i>embeddings</i> para NER	124
Figura 31 – Transferências programadas pela União Federal para saneamento básico:	
β estimado para municípios com mais de 20.000 habitantes	126
Figura 32 – Transferências programadas pela União Federal para saneamento básico:	
β estimado para municípios com mais de 50.000 habitantes	126
Figura 33 – Valores empenhados pela União Federal para saneamento básico: β	
estimado para municípios com mais de 20.000 habitantes	127
Figura 34 – Valores empenhados pela União Federal para saneamento básico: β	
estimado para municípios com mais de 50.000 habitantes	127

Lista de quadros

Quadro 1 – Questões urbanas, suas escalas temporal e espacial e o caráter das métricas associadas	67
Quadro 2 – Conjunto de etiquetas do Mac-Morpho	90
Quadro 3 – Sentenças do DOU com anotações morfossintáticas	120
Quadro 4 – Etiquetas previstas pelo modelo BLSTM	122

Lista de tabelas

Tabela 1 – Investimentos realizados em 2017, de acordo com as informações dos prestadores de serviços participantes do SNIS, segundo região geográfica e origem dos recursos (R\$ milhões)	26
Tabela 2 – Distribuição dos recursos federais não onerosos na LOA 2016, modalidade saneamento básico	27
Tabela 3 – Operações do FGTS para mutuários públicos - 2017	29
Tabela 4 – Operações do FGTS para o setor privado - 2017	30
Tabela 5 – Municípios metropolitanos e população, segundo faixas populacionais - 2016	72
Tabela 6 – RMs, população, porcentagem e número de municípios por RM, segundo faixas populacionais - 2016	74
Tabela 7 – Transferências da União Federal para saneamento básico por municípios selecionados - 2011 a 2018 (em R\$1,00)	85

Lista de abreviaturas e siglas

ANA	Agência Nacional de Águas
BNDES	Banco Nacional de Desenvolvimento Econômico e Social
CF	Constituição Federal
CEF	Caixa Econômica Federal
DOU	Diário Oficial da União
FAT	Fundo de Amparo ao Trabalhador
FGTS	Fundo de Garantia por Tempo de Serviço
FUNASA	Fundação Nacional de Saúde
GPS	Sistema de Geoposicionamento Global
IBGE	Instituto Brasileiro de Geografia e Estatística
LDNSB	Lei de Diretrizes Nacionais para o Saneamento Básico
LOA	Lei Orçamentária Anual
MP	Medida Provisória
OGU	Orçamento Geral da União
PAC	Programa de Aceleração do Crescimento
PIB	Produto Interno Bruto
PLANSAB	Plano Nacional de Saneamento Básico
PLN	Processamento de Linguagem Natural
PPA	Plano Plurianual
RM	Região Metropolitana
SIDOR	Sistema Integrado de Dados Orçamentários
SIGPLAN	Sistema de Informações Gerenciais de Planejamento do Plano Plurianual

SNIS	Sistema Nacional de Informações sobre Saneamento Básico
SIOP	Sistema Integrado de Planejamento e Orçamento do Governo Federal
TCU	Tribunal de Contas da União
TIC	Tecnologias de Informação e Comunicação
UO	Unidade Orçamentária
API	Application Program Interface
BLSTM	Bidirectional Long Short-Term Memory
CRF	Conditional Random Fields
HMM	Hidden Markov Model
HTML	HyperText Markup Language
LSTM	Long Short-Term Memory
MEM	Maximum Entropy Markov Model
MLP	MultiLayer Perceptron
NER	Named-entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NP	Noun Phrases
PDF	Portable Document Format
POS Tagger	Part-of-Speech Tagger
RNN	Recurrent Neural Network
SRN	Simple Recurrent Network
SQL	Structured Query Language
TnT	Statistical Part-of-Speech Tagger

Sumário

1	Introdução	17
2	Problema de pesquisa	20
2.1	<i>Marco legal do saneamento básico</i>	20
2.2	<i>Operação e financiamento do saneamento básico</i>	24
2.3	<i>Gastos federais com saneamento básico</i>	26
2.4	<i>Modalidades e mecanismos de transferência de recursos</i>	30
2.5	<i>Diagnóstico do saneamento básico</i>	32
2.6	<i>Síntese do problema de pesquisa</i>	39
2.7	<i>Objetivos da Pesquisa</i>	41
2.7.1	Objetivo geral	41
2.7.2	Objetivos específicos	42
2.7.3	Questões propostas	42
3	Referencial teórico	43
3.1	<i>Cidades como sistemas complexos</i>	43
3.1.1	Teoria geral de sistemas e cibernética	43
3.1.2	A nova ciência das cidades	48
3.1.3	Leis de escala em cidades	52
3.1.4	Modelos de redes aplicados a cidades	54
3.2	<i>Big data e planejamento urbano</i>	63
4	Metodologia da pesquisa	71
4.1	<i>Municípios e regiões metropolitanas</i>	71
4.2	<i>Plano de dados numéricos e textuais</i>	75
4.2.1	Estrutura e acesso ao DOU	75
4.2.2	Estrutura do SIOP	78
4.2.3	Extração de dados do SIOP	82
4.3	<i>Processamento de linguagem natural</i>	86
4.3.1	Pré-processamento de dados textuais	87
4.3.2	Etiquetagem morfosintática	88

4.3.3	Reconhecimento de entidades nomeadas	106
4.4	<i>Identificação de padrões de escalamento</i>	109
5	Resultados obtidos	113
5.1	<i>Softwares para extrair dados do DOU</i>	113
5.1.1	Softwares baseados em regras gramaticais	113
5.1.2	Modelo BLSTM	117
5.1.3	Avaliação de acurácia dos modelos	123
5.2	<i>Padrões de escalamento observados nos dados do SIOP</i>	124
6	Conclusões	128
	Referências ¹	133
	Apêndice A – Softwares para extração de informações do DOU	139
	Apêndice B – Ajuste de uma lei de escala aos dados SIOP .	143
	Anexo A – Treinamento de Etiquetadores Morfossintáticos .	145

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Essa dissertação de mestrado está dividida em seis Capítulos: (i) Introdução; (ii) Problema de pesquisa; (iii) Referencial teórico; (iv) Metodologia da pesquisa; (v) Resultados obtidos; e (vi) Conclusões.

O **Problema de pesquisa** situa-se na área de gestão de políticas públicas de saneamento básico em ambiente urbano.

Desde o início do século XX, vários programas governamentais têm buscado ampliar o acesso ao saneamento básico no Brasil. Embora tenham sido bem sucedidos em ampliar o acesso às redes de água e à coleta de lixo, mais de 100 milhões de pessoas, um pouco menos da metade da população brasileira, continua sem conexão às redes de coleta de esgoto. O sistema de esgotamento sanitário ainda é espacialmente concentrado na região Sudeste e nas áreas mais urbanizadas do País. Como explicar este padrão espacialmente concentrado? Como universalizar o acesso às redes de esgoto?

Dado que foram observadas leis de escala no volume de infraestrutura em relação ao tamanho das cidades, o aumento da urbanização, verificado nas últimas décadas, deveria ter facilitado a universalização dos serviços de saneamento básico, inclusive o acesso à coleta de esgotos. Uma hipótese a ser testada é se os esperados ganhos de escala no volume de infraestrutura de saneamento básico, decorrentes do aumento da urbanização, têm se traduzido, no caso brasileiro, em redução do montante de investimentos programados, de forma proporcional ao tamanho das cidades. O que deveria ser feito analisando-se, por meio de ferramentas estatísticas, os gastos com saneamento básico nos 5.570 municípios brasileiros, computando-se recursos provenientes de todas as fontes, públicas e privadas.

Face à complexidade dessa tarefa, reduzimos o escopo da pesquisa de forma a abranger somente as transferências de recursos, da União Federal para os municípios, na modalidade saneamento básico, por meio de recursos orçamentários ou de empréstimos aos operadores do sistema. Esperamos futuramente estendê-la às demais fontes de recursos, estaduais, municipais e privadas.

Para analisar as transferências de recursos da União Federal para os municípios selecionamos duas fontes de dados. A primeira é o Diário Oficial da União (DOU), periódico do governo federal para publicação de atos oficiais. A segunda é o Sistema Integrado de Planejamento e Orçamento do Governo Federal (SIOP), sistema de informação mantido

pela União Federal, com interface amigável e acesso livre e gratuito para qualquer cidadão conectado à internet.

A partir do Problema de pesquisa, definimos os Objetivos, Geral e Específicos, e um conjunto de Questões propostas.

O **Objetivo Geral** é explicar a concentração espacial dos serviços de saneamento básico no Brasil, a partir da análise dos padrões de escalamento das transferências de recursos, da União Federal para os municípios brasileiros, na modalidade saneamento básico, em relação à população destes municípios. Partimos da premissa de que os ganhos de escala na infraestrutura de saneamento básico devidos à urbanização deveriam resultar em decréscimo nos custos, de forma proporcional ao tamanho da cidade, levando a uma (esperada) relação de lei de potência (ou de escala) entre os gastos com saneamento básico e o tamanho da cidade. Se o decréscimo nos gastos, proporcional ao tamanho dos municípios, não for verificado na prática, pode-se concluir que os municípios maiores e mais ricos estão recebendo mais recursos do que o necessário, em detrimento dos demais, o que explicaria, em grande medida, a concentração espacial dos serviços de saneamento básico.

Como a extração de dados do DOU requer ferramentas de mineração de dados textuais, o **primeiro Objetivo Específico** é desenvolver um *software* para tal fim. O **segundo Objetivo Específico** é identificar os padrões de escalamento das transferências de recursos para saneamento básico, da União Federal para os municípios, em relação ao seu tamanho, medido em termos populacionais.

O **Referencial teórico** está dividido em três Seções. A primeira abrange os principais aspectos da Teoria Geral de Sistemas e da Cibernética aplicáveis ao planejamento e gestão de cidades, e as mudanças recentes nesse paradigma, introduzidas pelas novas Ciências da Complexidade. A segunda abrange as leis de escala e modelos de crescimento de redes capazes de explicar as morfologias urbanas. Na terceira analisamos os possíveis usos de grandes volumes de dados nos projetos de cidades inteligentes, possibilidades e limitações da aplicação desses dados ao planejamento e gestão de cidades.

A **Metodologia** está dividida em quatro Seções. Inicialmente tentamos explicar o porquê da escolha dos municípios, em vez das Regiões Metropolitanas, como equivalentes das cidades, tal como definidas no Referencial Teórico. A seguir vem o Plano de Dados Numéricos e Textuais, contendo descrição, o mais detalhada possível, da estrutura das fontes e do processo de extração dos dados. A Seção seguinte contempla as ferramentas de NLP aplicáveis à extração de dados do DOU: algoritmos de aprendizado de máquina (*machine*

learning) e Redes Bidirecionais de Memória de Curto e Longo Prazos (*Bidirectional Long Short-Term Memory - BLSTM*). A última Seção é dedicada às ferramentas estatísticas utilizadas na detecção de padrões de escalamento nas transferências da União Federal para os municípios.

O Capítulo de **Resultados obtidos** mostra os *softwares* desenvolvidos para a extração de dados do DOU e os padrões de escalamento observados nos dados do SIOP.

Nas **Conclusões** buscamos responder às Questões propostas.

O **Anexo A** contém os códigos Python correspondentes ao treinamento de diversos etiquetadores morfossintáticos e do modelo BLSTM, mostrados na Metodologia.

Finalmente, os dois *softwares*, correspondentes ao primeiro Objetivo Específico, estão no **Apêndice A**. O código Python para ajuste de uma lei de escala aos dados SIOP, correspondente ao segundo Objetivo Específico, está no **Apêndice B**.

2 Problema de pesquisa

2.1 *Marco legal do saneamento básico*

Conforme disposto na Constituição Federal (CF) de 1998, sobre as competências dos três entes federativos, a União Federal tem competência para estabelecer as diretrizes gerais do setor. Tanto a União Federal quanto os estados e os municípios devem promover programas de melhoria do saneamento básico. Compete ainda aos municípios legislar sobre os assuntos e organizar e prestar os serviços de interesse local (BRASIL, 1998).

O marco legal do saneamento básico é composto pela Lei nº 11.445/2007, denominada Lei de Diretrizes Nacionais para o Saneamento Básico (LDNSB), regulamentada pelo Decreto nº 7.217/2010.

A LDNSB estabelece o leque de serviços que compõem o saneamento básico: abastecimento de água potável, esgotamento sanitário, limpeza urbana e manejo de resíduos sólidos, além de drenagem urbana e manejo de águas pluviais. Há ainda legislação específica para a gestão de resíduos sólidos, Lei nº 12.305/2010, conhecida como Política Nacional de Resíduos Sólidos, regulamentada pelo Decreto nº 7.404/2010 (BRASIL, 2007; BRASIL, 2010a; BRASIL, 2010c; BRASIL, 2010b).

A LDNSB também delimita a abrangência dos componentes do saneamento básico. O abastecimento de água potável é constituído pelas atividades, infraestruturas e instalações necessárias ao abastecimento público de água potável, desde a captação até as ligações prediais e respectivos instrumentos de medição. O esgotamento sanitário é constituído pelas atividades, infraestruturas e instalações operacionais de coleta, transporte, tratamento e disposição final adequados dos esgotos sanitários, desde as ligações prediais até o seu lançamento final no meio ambiente. A limpeza urbana e o manejo de resíduos sólidos são definidos como um conjunto de atividades, infraestruturas, e instalações operacionais de coleta, transporte e transbordo, tratamento e destino final do lixo doméstico e do lixo originário da varrição e limpeza de logradouros e vias públicas. Por fim, a drenagem e o manejo das águas pluviais urbanas são definidos como um conjunto de atividades, infraestruturas, e instalações operacionais de drenagem urbana de águas pluviais, de transporte, detenção ou retenção para o amortecimento de vazões de cheias, tratamento e disposição final das águas pluviais drenadas nas áreas urbanas (BRASIL, 2007).

A titularidade dos serviços de saneamento básico pertence aos municípios e ao Distrito Federal. Quanto às Regiões Metropolitanas (RMs), essa titularidade não está definida de forma consolidada, em razão de disputas judiciais entre estados e municípios. Para solucionar problemas que afetem mais de um município é permitido às administrações municipais se consorciarem para implantar e executar projetos e ações de maior porte ou de abrangência regional, dividindo os custos e compartilhando os benefícios, por meio dos assim denominados consórcios municipais.

Para ilustrar a disputa judicial sobre a titularidade dos serviços de saneamento, recordemos que, no caso do município de São Paulo, até 1870, os serviços de saneamento básico eram prestados diretamente pelo poder público local, quando se fez a concessão a um operador privado, a Companhia de Água e Esgotos da Cantareira. Em 1892, o Estado de São Paulo encampou a operação dos serviços e passou a prestá-los diretamente. Após sucessivas reorganizações administrativas, a SABESP, criada em 1973, passou a operar a rede, sem celebrar qualquer contrato de concessão com o município de São Paulo (CUNHA, 2011).

A defesa da SABESP mobilizou recursos jurídicos e políticos, que produziram teses em favor da titularidade estadual sobre os serviços de saneamento básico. A mais comum entre essas teses sustenta que a titularidade deve ser estadual sempre que a montagem da cadeia industrial necessária à operação ultrapasse as fronteiras de um único município, o que é frequente nas RM. Essa questão não está totalmente resolvida do ponto de vista jurídico, a despeito do disposto no arcabouço regulatório em vigor. Porém, ao celebrar, em 2009, um contrato de concessão com o município de São Paulo, a SABESP passou a reconhecer tacitamente a titularidade municipal (CUNHA, 2011).

A LDNSB determinou à União Federal a elaboração do Plano Nacional de Saneamento Básico (PLANSAB), concluído em 2013. Conforme consta no PLANSAB, o montante de investimentos necessário para universalizar os serviços de saneamento básico é de R\$ 508 bilhões, a ser realizado entre 2013 e 2033, por meio de metas nacionais e regionalizadas de curto, médio e longo prazos. Somente para a universalização dos serviços de água e esgotamento sanitário no Brasil estima-se que seja necessário investir R\$ 15,6 bilhões anuais até 2033 (BRASIL, 2014).

Segundo a referida Lei, cabe aos municípios a elaboração de planos de saneamento básico, o estabelecimento de sistemas de informações sobre os serviços e a instituição de mecanismos de controle social. Para ter acesso aos recursos da União, os municípios

precisam elaborar Planos Municipais de Saneamento Básico e instituir, por meio de lei municipal, um órgão colegiado para efetivar o controle social. O controle social é definido como “um conjunto de mecanismos e procedimentos que garantem a sociedade informações, representações técnicas e participações nos processos de formulação de políticas, de planejamento e de avaliação relacionados aos serviços públicos de saneamento básico” (BRASIL, 2007).

A existência de plano de saneamento básico condiciona o acesso aos recursos orçamentários da União Federal, ou por ela administrados, quando destinados a serviços de saneamento básico, como estabelece o Artigo 26 do Decreto 7.217/2010. Neste mesmo decreto, a Presidência da República determina que os municípios que se interessassem em pleitear recursos da União deveriam elaborar o seu plano de saneamento básico até 31 de dezembro de 2017, contando com o órgão colegiado já instituído para exercer o controle social (BRASIL, 2017).

Levantamento divulgado, em janeiro de 2017, pelo extinto Ministério das Cidades, mostrava que apenas 30% do total de 5.570 municípios brasileiros tinham concluído os seus planos municipais e outros 38% estariam em fase de elaboração do plano (figura 1). Em suma, 68% dos municípios brasileiros (3.785) declararam ter elaborado ou estarem em fase de elaboração dos respectivos planos. Destes municípios, a grande maioria (89,7%) tem população inferior a 50 mil habitantes. No grupo de municípios com população superior a 50 mil habitantes, apenas 388 municípios declararam possuir um plano elaborado ou em fase de elaboração (BRASIL, 2017).

Na perspectiva dos usuários dos serviços de saneamento urbano, o aspecto mais relevante da LDNSB é a obrigatoriedade de conexão das edificações permanentes urbanas às redes públicas de abastecimento de água e de esgotamento sanitário disponíveis, sujeita ao pagamento das tarifas e de outros preços públicos decorrentes da conexão e do uso desses serviços. Na ausência de redes públicas de saneamento básico, são admitidas soluções individuais de abastecimento de água e de afastamento e destinação final dos esgotos sanitários, conforme normas regulatórias. Para reforçar a obrigatoriedade de conexão à rede, governos locais estão impondo multas a quem não faz a ligação. Em algumas prefeituras, se a rede passa na frente da residência, o usuário paga a taxa de serviços estando ou não conectado (BRASIL, 2007).

Figura 1 – Panorama dos planos municipais de saneamento básico - 2017

Fonte: Brasil (2017)

Mudanças no marco legal do saneamento básico foram propostas por meio da Medida Provisória (MP) nº 844/2018, que, dentre outros dispositivos, atribuiu à Agência Nacional de Águas (ANA) a competência para editar normas de referências nacionais sobre o serviço de saneamento. As normas de referência seriam instituídas pela ANA de forma progressiva. O acesso aos recursos federais, destinados aos serviços de saneamento básico, estariam condicionados ao cumprimento das normas de referência estabelecidas pela ANA. A MP 844/2018 não chegou a ser apreciada pelo Congresso Nacional e perdeu a validade em novembro do mesmo ano (BRASIL, 2018d).

A MP 868/2018 incorporou parte do que previa a MP 844/2018 e definiu como atribuições da ANA fazer estudos de mercado e contratar agentes para prestação de serviços públicos. Nesse último caso, a ANA deverá definir metas de qualidade, eficiência e ampliação da cobertura de serviços. Ela determina que os contratos de saneamento passarão a ser estabelecidos por meio de processo licitatório com chamamento público “para verificar o interesse dos entes federativos, em regime isolado ou consorciado, em realizar

concessões e parcerias público-privadas”. Antes, a licitação poderia ser dispensada caso o contrato fosse firmado com uma empresa pública. Caso seja aprovada, a regulamentação dos serviços de águas e esgotos, que é de atribuição dos municípios, passará a ser feita pela agência reguladora, também responsável pela fixação das tarifas (BRASIL, 2018e).

2.2 Operação e financiamento do saneamento básico

A operação dos serviços de saneamento básico abrange a provisão e manutenção da cadeia industrial e a oferta efetiva dos serviços públicos. O operador pode ser o próprio titular ou um terceiro. A prestação direta pode ocorrer por meio da própria administração pública direta ou de autarquias, fundações públicas, empresas públicas ou de economia mista vinculadas ao ente federado que detém a titularidade. Todavia a LDNSB permite que o titular entregue a operação dos serviços de saneamento básico a terceiros, por meio de um processo licitatório e da consequente celebração de um contrato de concessão com um concessionário, público ou privado (CUNHA, 2011).

Em virtude do seu caráter de cadeia industrial e rede de serviços públicos, é possível desmembrar e delegar cada etapa da operação dos serviços de saneamento básico a um operador diferente. Por exemplo, quem presta serviços de fornecimento de água não necessariamente deve prover os serviços de esgoto. Mesmo a operação da rede de distribuição de água pode dar-se independentemente do controle das atividades de captação, tratamento e adução da água bruta (CUNHA, 2011).

As principais fontes de investimento em saneamento básico no Brasil são: (i) recursos onerosos, provenientes do Fundo de Garantia por Tempo de Serviço (FGTS) e do Fundo de Amparo ao Trabalhador (FAT), sob gestão da União Federal, e também derivados de empréstimos internacionais, obtidos junto às agências multilaterais de crédito; (ii) recursos não onerosos, oriundos do Orçamento Geral da União (OGU), constantes na Lei Orçamentária Anual (LOA), e nos orçamentos dos estados e municípios, que não preveem retorno financeiro, pois os agentes beneficiados não precisam ressarcir os cofres públicos; e (iii) recursos próprios dos prestadores de serviços, resultantes da cobrança por esses serviços (SANEAMENTO-SNIS, 2017b).

Conforme glossário do Sistema Nacional de Informações sobre Saneamento Básico (SNIS), os investimentos realizados com recursos próprios são aqueles cujas obras são

pagas com recursos do caixa do prestador de serviços, resultante da cobrança de tarifa; os investimentos com recursos onerosos são aqueles cujas obras são pagas com recursos de financiamentos, retornáveis por meio de amortizações, juros e encargos. E os investimentos com recursos não onerosos são os não reembolsáveis, que não oneram o serviço da dívida, também denominados recursos a fundo perdido (SANEAMENTO-SNIS, 2017b).

Os valores de investimentos do SNIS correspondem a montantes efetivamente realizados no ano, informados pelos prestadores de serviços de água e esgotos. Essa categoria não deve ser confundida com outras formas de contabilização de investimentos, tais como: recursos empenhados, recursos desembolsados ou recursos contratados, adotados pelos sistemas federais de controle de investimentos (SANEAMENTO-SNIS, 2017b).

Em 2017, a totalização dos investimentos, segundo a fonte dos recursos, resultou em R\$ 6.327,2 (58,7%) realizados com recursos próprios, R\$ 2.993,4 (27,2%) realizados com recursos onerosos e R\$ 1.515,2 milhões (14,1%) com recursos não onerosos. No total, foram realizados investimentos iguais a R\$ 10.775,8 milhões (SANEAMENTO-SNIS, 2017b).

Na comparação com 2016, a participação relativa dos investimentos realizados com recursos próprios cresceu de 55,7%, em 2016, para 58,7% em 2017. Em valores absolutos, os investimentos realizados com recursos próprios aumentaram de R\$ 6.306,7 milhões, em 2016, para R\$ 6.327,2 milhões, em 2017 (SANEAMENTO-SNIS, 2017b).

No documento "Diagnóstico dos Serviços de Água e Esgoto 2017" constam ressalvas sobre a elevada participação relativa dos recursos próprios. Ela seria uma indicação de que os prestadores de serviços estariam declarando como próprios, investimentos feitos com recursos dos governos dos estados e das prefeituras municipais, ou até mesmo recursos oriundos de transferências da União Federal, previstos na LOA. Poderia também acontecer de alguns prestadores de serviços classificarem como próprios os recursos oriundos de empréstimos por entenderem que, uma vez que os mesmos deverão ser necessariamente pagos, então é possível o entendimento de que seriam recursos próprios (SANEAMENTO-SNIS, 2017b).

A tabela 1 mostra a divisão de recursos por região e origem dos mesmos.

Tabela 1 – Investimentos realizados em 2017, de acordo com as informações dos prestadores de serviços participantes do SNIS, segundo região geográfica e origem dos recursos (R\$ milhões)

Região	Recursos Próprios	Onerosos	Não-Onerosos	Total	%
Norte	152,8	149,9	153,8	456,6	4,2
Nordeste	788,1	227,7	852,0	1.867,9	17,3
Sudeste	3.663,1	1.810,9	194,2	5.668,2	52,6
Sul	1.394,0	420,8	71,8	1.886,6	17,5
Centro-Oeste	329,1	324,0	243,4	896,6	8,3
Brasil	6.327,2	2.993,4	1.515,2	10.775,8	100,0
%	58,7%	27,2%	14,1%	100,0%	-

Fonte: SANEAMENTO-SNIS (2017b)

Quanto aos investimentos com recursos próprios, a região Sudeste apresenta valores bem superiores às demais, com R\$ 3.663,1 milhões, 52,6% do total de recursos desta fonte. Por outro lado, a região Norte apresenta o menor valor, equivalente a R\$ 152,8 milhões, 2,4% do total de recursos próprios investidos.

Com relação aos recursos onerosos, a região Sudeste desponta com valores bem superiores às demais, declarando ter investido, em 2017, R\$ 1.810,9 milhões, 60,4% do total desta fonte, enquanto a região Norte novamente aparece com o menor valor investido, igual a R\$ 149,9 milhões, cerca de 5% do total de recursos onerosos.

Por fim, quanto aos investimentos com recursos não onerosos, a região Nordeste apresentou o maior valor, igual a R\$ 852,0 milhões, aproximadamente 56,2% do total desta fonte, enquanto que a região Sul, com R\$ 71,8 milhões, correspondente a cerca de 4,7% do total de recursos não onerosos, aplicou o menor montante (SANEAMENTO-SNIS, 2017b).

2.3 Gastos federais com saneamento básico

Nesse item analisaremos os gastos da União Federal com saneamento básico, onerosos e não onerosos. O último relatório consolidado disponível sobre os gastos federais, produzido pelo extinto Ministério das Cidades, refere-se ao ano de 2016. Quanto ao FGTS, o último relatório de gestão disponível, produzido pelo extinto Ministério do Trabalho, refere-se ao ano de 2017. O último relatório de gestão do FAT também se refere ao ano de 2017. Esse último não contém uma descrição desagregada, por estado ou por região, da execução dos recursos do FAT destinados ao saneamento básico, a cargo do Banco Nacional de

Desenvolvimento Econômico e Social (BNDES) por isso não será utilizado como fonte de informações nesse item (BRASIL, 2018b; BRASIL, 2018g; BRASIL, 2018f)

Se considerarmos apenas os recursos não onerosos, previstos na LOA, foram destinados, em 2016, R\$ 3,66 bilhões para saneamento básico. O extinto Ministério das Cidades foi o órgão contemplado com a maior parcela de recursos, 67,8% do total programado para o período (tabela 2). Ao Ministério da Saúde, ao qual se vincula a Fundação Nacional de Saúde (FUNASA), foram destinados 20,5% do total. Os restantes 11,7% foram distribuídos entre os atualmente extintos Ministérios da Integração Nacional e Desenvolvimento Social, e ao Ministério do Meio Ambiente (BRASIL, 2018b).

Tabela 2 – Distribuição dos recursos federais não onerosos na LOA 2016, modalidade saneamento básico

Órgão	Recursos não onerosos (R\$1,00)	%
Min. Cidades	2.478.756.787	67,8
Min. Saúde	748.768.743	20,5
Min. Integração Nacional	269.963.783	7,4
Min. Desenvolvimento Social	154.926.719	4,2
Min. Meio Ambiente	3.914.205	0,1

Fonte: Brasil (2018b)

Os investimentos previstos na LOA podem ser livres ou provenientes de emendas parlamentares. Os recursos livres estão sob a governabilidade dos órgãos setoriais. Na LOA 2016, foi destinado R\$ 1,70 bilhão em recursos para investimentos de livre alocação (46,5% do total). Deste montante, R\$ 1,43 bilhão (83,9%) são recursos programados para investimentos do Programa de Aceleração do Crescimento (PAC) (BRASIL, 2018b).

Por sua vez, as emendas parlamentares têm vinculação específica e destinatário determinado pelos parlamentares. Na LOA 2016, do montante de recursos aprovados para investimentos em saneamento básico (R\$ 3,66 bilhões), 53,5%, ou R\$ 1,96 bilhão, foram decorrentes de emendas parlamentares. De 2004 a 2016, 43,6%, em média, dos recursos programados foram destinados a emendas tal como nos mostra a figura 2 (BRASIL, 2018b).

Considerando-se apenas as operações realizadas em 2016, a cargo do extinto Ministério das Cidades, 74,7% do número total de contratos corresponderam a investimentos com recursos não onerosos. No entanto, o valor médio de investimento por operação de empréstimo, com recursos FGTS e FAT, foi cerca de quatro vezes superior às operações contratadas com recursos não onerosos. Enquanto o valor médio dos investimentos com recursos orçamentários foi de R\$ 12,9 milhões, o valor médio dos investimentos com

recursos de empréstimos foi de R\$ 50,9 milhões. Em volume de recursos, nada menos que 57,2% dos investimentos em saneamento básico do Ministério das Cidades foram oriundos de contratos de empréstimos, com predominância nesse conjunto dos recursos do FGTS(69,9%) (BRASIL, 2018b).

Figura 2 – Evolução dos recursos da LOA e percentual de emendas parlamentares - 2004 a 2016

* As séries foram indexadas pelo IGP-DI médio anual da FGV para o ano de 2016

Fonte: Brasil (2018b)

Na análise dos empréstimos totais, por fonte e por região, 64,1% dos recursos foram destinados, em 2016, à Região Sudeste, que abriga os municípios mais ricos e de maior porte, e 48,0% dos desembolsos de financiamentos com recursos do FAT foram destinados à mesma região. A causa da concentração regional teria sido a dificuldade de acesso ao crédito pelos municípios das demais regiões (BRASIL, 2018b).

Em 2017, cerca de 30% do volume de recursos destinados pelo FGTS aos mutuários públicos e cerca de 33% dos recursos destinados aos operadores privados, para saneamento básico, foram contratados pelo estado de São Paulo. Mais de 80% dos recursos contratados pelo setor privado foram destinados ao esgotamento sanitário.

Detalhando a atuação do FGTS em 2017, na área de saneamento básico, temos a seguinte situação. Para os mutuários públicos, foram habilitadas e contratadas 131 operações de crédito no valor de R\$ 3.789.287.555,26, em oito estados, para as modalidades de esgotamento sanitário, abastecimento de água, desenvolvimento institucional, redução e controle de perdas, e estudos e projetos, conforme detalhado na tabela 3. Cerca de 30% do volume de recursos destinados, em 2017, pelo FGTS, aos mutuários públicos, na modalidade saneamento básico, foram contratados por São Paulo, estado da Região Sudeste. Quanto às operações do setor privado, mostradas na tabela 4, foram enquadradas 32 propostas em nove estados brasileiros, nas modalidades de abastecimento de água, esgotamento sanitário, redução e controle de perdas, manejo de resíduos sólidos e desenvolvimento institucional, totalizando um valor de aproximadamente R\$ 2 bilhões (BRASIL, 2018g).

Tabela 3 – Operações do FGTS para mutuários públicos - 2017

UF/Modalidade	Quantidade	Valor do empréstimo(R\$)
ES	5	137.122.533,60
Abastecimento de água	1	73.557.715,11
Desenvolvimento institucional	1	14.197.750,00
Esgotamento sanitário	3	49.367.068,49
GO	2	341.386.571,10
Abastecimento de água	2	41.386.571,10
MS	1	4.108.910,00
Redução e controle de perdas	1	4.108.910,00
PE	9	312.490.000,00
Abastecimento de água	8	276.390.000,00
Esgotamento sanitário	1	36.100.000,00
PR	102	1.826.197.870,03
Abastecimento de água	41	697.578.985,84
Desenvolvimento institucional	5	147.277.330,00
Esgotamento sanitário	54	949.270.971,01
Estudos e projetos	1	4.507.750,00
Redução e controle de perdas	1	27.562.833,18
RJ	1	570.086.937,24
Abastecimento de água	1	570.086.937,24
RS	9	210.568.952,19
Abastecimento de água	1	6.368.542,62
Esgotamento sanitário	8	204.200.409,57
SP	2	387.325.781,10
Abastecimento de água	1	127.577.459,79
Esgotamento sanitário	1	259.748.321,31
TOTAL	131	3.789.287.555,26

Fonte: Brasil (2018g)

Tabela 4 – Operações do FGTS para o setor privado - 2017

UF	Municípios beneficiados	Quantidade	VE (R\$)
AL	Maceió	01	140.740.542,00
MA	Timon	03	79.495.688,00
MT	Confresa	03	29.450.974,00
MT	Diamantino	03	18.313.574,00
PE	RM Recife e Goiana	01	634.475.445,00
RO	Buritis	01	634.475.445,00
RS	Uruguaiana	01	47.310.000,00
SC	Balneário Camboriú	02	60.641.312,00
SC	Penha	03	68.617.911,00
SC	São Francisco do Sul	03	75.049.092,00
SP	Americana	01	3.583.400,00
SP	Caieiras	01	73.374.165,00
SP	Mairinque	01	41.480.069,00
SP	Ribeirão Preto	01	130.851.489,00
SP	Santa Gertrudes	01	4.003.929,00
SP	Sumaré	03	401.909.049,00
TO	Colinas do Tocantins	01	20.300.185,00
TO	Guaraí	01	3.711.477,00
TO	Palmas	01	128.856.852,00
TOTAL	-	32	1.987.029.759,00

Fonte: Brasil (2018g)

A distribuição dos recursos contratados em 2018 mostra que os mesmos foram destinados a apenas cinco estados: Minas Gerais (35% dos recursos); São Paulo (24%), Mato Grosso do Sul(23%), Paraná(17%) e Santa Catarina(1%). Ou seja, em 2018, assim como em 2016, houve uma grande concentração dos recursos do FGTS na Região Sudeste, da ordem de 60%.

2.4 Modalidades e mecanismos de transferência de recursos

Até 2018, a maior parte das transferências de recursos da União Federal para os municípios, na modalidade saneamento básico, estava a cargo do atualmente extinto Ministério das Cidades e da FUNASA, fundação vinculada ao Ministério da Saúde. Informações disponíveis no sítio da internet do extinto Ministério das Cidades mencionavam uma repartição de competências, estabelecida na esfera federal, sobre o repasse de recursos, descrita a seguir.

Quanto ao abastecimento de água, esgotamento sanitário e manejo de resíduos sólidos urbanos, cabia ao Ministério das Cidades o atendimento a municípios com população

superior a 50.000 habitantes ou integrantes de Regiões Metropolitanas, Regiões Integradas de Desenvolvimento ou participantes de Consórcios Públicos afins. Para os municípios com população inferior a 50.000 habitantes, o Ministério das Cidades só atuava por meio de financiamento com recursos onerosos para as modalidades de abastecimento de água e esgotamento sanitário.

Para os municípios com população de até 50.000 habitantes, o atendimento com recursos não onerosos era realizado pelo Ministério da Saúde, por meio da FUNASA. Particularmente, com relação ao manejo de águas pluviais urbanas, a competência era compartilhada entre os atualmente extintos Ministérios das Cidades e da Integração Nacional, além de intervenções da FUNASA em áreas com forte incidência de malária.

O Ministério das Cidades participava da gestão dos seguintes programas temáticos do Plano Plurianual (PPA) do governo federal: Programa Saneamento Básico (2068); Programa Planejamento Urbano (2054); e Programa de Gestão de Riscos e Resposta a Desastres (2040). O PPA é um instrumento previsto no Artigo 165 da CF destinado a organizar e viabilizar a ação pública. Por meio dele, é declarado o conjunto das políticas públicas do governo para um período de quatro anos e as ações para viabilizar as metas previstas (ENAP, 2015b; ENAP, 2015a).

Também estavam sob a supervisão do Ministério das Cidades, o PAC e o Programa Saneamento para Todos, que não são Programas Temáticos do PPA. Na realidade, as ações e os recursos do PAC estavam dispersos em vários Programas Temáticos do PPA, inclusive nos três programas citados. O Programa Saneamento para Todos, embora também tivesse recursos dispersos em vários Programas Temáticos do PPA, era um programa que utilizava recursos de natureza extraorçamentária, mais especificamente do FGTS (ENAP, 2015b; ENAP, 2015a).

Além dos recursos do Programa Saneamento Para Todos, oriundos do FGTS, o Ministério das Cidades também utilizava, em seus processos seletivos de operações de crédito, recursos oriundos de outras fontes extraorçamentárias: recursos geridos pelo BNDES, próprios ou do FAT. Em suma, as ações do Ministério das Cidades contavam com dois tipos de fontes de recursos: (i) recursos não onerosos: constantes na LOA; e (ii) recursos onerosos: de natureza extraorçamentária, oriundos do FGTS, FAT e de recursos próprios de outros agentes financeiros (ENAP, 2015b; ENAP, 2015a).

O acesso aos recursos da LOA dependia do tipo de dotação orçamentária: (i) as dotações consignadas no orçamento do Ministério das Cidades se submetiam a processos

seletivos; e (ii) as dotações decorrentes de emendas parlamentares eram dispensadas de seleção, por sua natureza individualizada de alocação do gasto, definida na própria lei orçamentária.

O acesso aos recursos extraorçamentários era até então regulamentado, por leis e decretos federais, por Instruções Normativas editadas pelo Ministério das Cidades e por regulamentação publicada por outros órgãos e entidades, dentre os quais o Conselho Monetário Nacional e o Conselho Curador do FGTS.

Acrescente-se que o instrumento jurídico que formalizava o repasse de recursos nos projetos do PAC era o Termo de Compromisso, enquanto que nos projetos que não eram PAC o instrumento era o Convênio, quando celebrado diretamente pelo Ministério. Por sua vez, o Contrato de Repasse era operacionalizado por instituição financeira, por exemplo, a Caixa Econômica Federal (CEF), contratada especificamente para esse fim (ENAP, 2015a).

Finalmente, segundo informações disponíveis no sítio da FUNASA na internet, as ações de saneamento desenvolvidas por essa fundação eram custeadas com recursos não onerosos, e executadas por meio de Convênios celebrados diretamente com os municípios e estados. Em casos excepcionais, a FUNASA executava direta ou indiretamente as ações.

Os sistemas de informação da esfera federal não permitem a extração direta de todas as informações sobre as transferências da União Federal para municípios, na modalidade saneamento básico. A destinação dos recursos não onerosos, previstos na LOA, pode ser consultada livremente por meio do SIOP.

2.5 *Diagnóstico do saneamento básico*

As informações sobre a situação do saneamento básico no Brasil estão reunidas no SNIS, o qual se encontrava, até janeiro de 2019, sob a responsabilidade do Ministério das Cidades, assumida a partir de então pelo Ministério do Desenvolvimento Regional, resultante da fusão dos extintos Ministérios das Cidades e da Integração Nacional.

Com base nas informações do SNIS, são elaborados anualmente dois documentos: "Diagnóstico de Serviços de Água e Esgoto"; e "Diagnóstico do Manejo de Resíduos Sólidos Urbanos". As versões mais recentes destes documentos referem-se ao ano de 2017.

Sobre os serviços de água e de esgotos, os dados são atualizados anualmente para uma amostra de prestadores de serviços do Brasil, desde o ano-base de 1995. Quanto

aos serviços de manejo de resíduos sólidos, os dados são também atualizados anualmente para uma amostra de municípios brasileiros, desde o ano-base de 2002. Os dados dos serviços de água e esgoto são fornecidos ao SNIS por companhias estaduais, empresas e autarquias municipais, empresas privadas e, em muitos casos, pelas próprias prefeituras, todos denominados no SNIS como prestadores de serviços. No caso do manejo de resíduos sólidos, as informações são provenientes de órgãos gestores dos serviços nos municípios, quais sejam, autarquias, departamentos ou secretarias municipais (SANEAMENTO-SNIS, 2017b; SANEAMENTO-SNIS, 2017a).

Em 2017, o SNIS apurou informações sobre abastecimento de água em 5.126 municípios, com população urbana de 172,1 milhões de habitantes, assegurando uma representatividade de 92,0% em relação ao total de municípios e de 98,0% em relação à população urbana do Brasil. Para o esgotamento sanitário, a quantidade de municípios foi de 3.865 e a população urbana de 161,4 milhões de habitantes, uma representatividade de 69,4% em relação ao total de municípios e de 91,9% em relação à população urbana do Brasil (SANEAMENTO-SNIS, 2017b).

Segundo o SNIS, o contingente de população urbana atendida por redes de esgoto, em 2017, alcançava 105,3 milhões de habitantes, um incremento de 1,4 milhão de novos habitantes atendidos, crescimento de 1,3%, na comparação com 2016. Já o índice médio de atendimento era de 60,2% nas áreas urbanas das cidades brasileiras, destacando-se a região Sudeste, com média de 83,2% (SANEAMENTO-SNIS, 2017b).

Mas a melhoria verificada nos índices de atendimento urbano, de 2017 em relação a 2016, foi insuficiente para alterar a situação sanitária de forma significativa. Em 2017, cerca de 49,7% dos habitantes, ou seja, um pouco menos da metade da população brasileira, continuava sem acesso a redes de coleta de esgoto. O que significa que mais de 100 milhões de pessoas utilizavam fossas sépticas ou rudimentares ou ainda descartavam o esgoto *in natura* em rios, lagos ou mar (SANEAMENTO-SNIS, 2017b).

Quanto ao tratamento dos esgotos, observa-se que o índice médio do País chegava a 46,0% para a estimativa dos esgotos gerados e 73,7% para os esgotos que são coletados. O volume de esgotos tratados passou de 4,1 bilhões de m^3 , em 2016, para 4,2 bilhões de m^3 , em 2017, um incremento de 3,0% (SANEAMENTO-SNIS, 2017b).

Por sua vez, o contingente de população urbana atendida por redes de água era igual a 159,9 milhões de habitantes, em 2017, o que representa um incremento de 1,1 milhão de novos habitantes atendidos, crescimento de 0,7%, na comparação com

2016. Quanto ao índice de atendimento, observam-se valores bastante elevados nas áreas urbanas das cidades brasileiras, com uma média nacional de 93,0%. Destacam-se as regiões Sul, Centro-Oeste e Sudeste, em que os índices médios são de 98,4%, 98,1% e 95,9%, respectivamente (SANEAMENTO-SNIS, 2017b).

Na figura 4 vemos a distribuição espacial do índice de atendimento urbano por redes coletoras de esgoto, a qual mostra uma concentração destes serviços nos grandes aglomerados urbanos do Sudeste e regiões contíguas. A figura 3 representa a distribuição espacial do índice de atendimento urbano por redes coletoras de água, a qual mostra um padrão mais difuso que o anterior, não mais concentrado na região Sudeste.

Quanto aos serviços de manejo de resíduos sólidos, em 2017, 3.556 municípios participaram da coleta, isto é, 63,8% do total. Em termos de população urbana este percentual equivale a 83,9% ou 146,3 milhões de habitantes.

Segundo dados do SNIS, a cobertura do serviço regular de coleta domiciliar era, em 2017, igual a 98,6% da população urbana, equivalente a um déficit de atendimento a aproximadamente 2,7 milhões de habitantes das cidades brasileiras, sendo 44,7% destes, moradores da região Nordeste, 21,8% da região Sudeste, 21,1% da região Norte e outros 12% divididos entre as regiões Sul e Centro-Oeste (SANEAMENTO-SNIS, 2017a).

O maior déficit de atendimento do serviço de coleta regular foi encontrado nos municípios com até 30.000 habitantes, muito embora seu indicador médio ainda permaneça com percentual elevado, de 97,4%. As taxas de cobertura do serviço de coleta de resíduos sólidos, em relação à população urbana nos municípios abrangidos pelo SNIS, estão mostradas na figura 5.

As faixas populacionais consideradas pelo SNIS-RS foram as seguintes :

- faixa 1: até 30 mil habitantes;
- faixa 2: de 30.001 a 100 mil habitantes;
- faixa 3: de 100.001 a 250 mil habitantes;
- faixa 4: de 250.001 a 1,0 milhão de habitantes;
- faixa 5: de 1.000.001 a 3,0 milhões de habitantes; e
- faixa 6: acima de 3,0 milhões de habitantes.

As diferenças nas taxas de urbanização parecem explicar, em grande medida, o padrão regional de acesso às redes de água e esgoto no Brasil.

A região Sudeste é a que apresenta as maiores taxas de urbanização. Dados do Instituto Brasileiro de Geografia e Estatística (IBGE), referentes a 2010, revelam uma taxa de urbanização da ordem de 92,95%. Além disso, esta região abriga os dois maiores centros urbanos do País: as cidades de São Paulo e Rio de Janeiro. Na região Centro-Oeste, o processo de urbanização teve como principal fator a construção de Brasília, inaugurada em 1960. Desde o final da década de 1960 e início da década de 1970, o Centro-Oeste tornou-se a segunda região mais urbanizada do País. Em 2010 a taxa de urbanização dessa região era de 88,8%. A urbanização na região Sul foi lenta até a década de 1970, mas se acelerou nas últimas décadas, chegando a 84,93% em 2010. Em contraposição, temos o Norte e o Nordeste do País, onde as taxas de urbanização eram de aproximadamente 73%, em 2010, também segundo dados do IBGE.

Por outro lado, na coleta de resíduos sólidos, as análises mostram que o tamanho dos municípios influencia a cobertura dos serviços.

Com base nesse diagnóstico, podemos inferir que as desigualdades na oferta de serviços de saneamento básico estejam relacionadas tanto às diferenças regionais, em termos da taxas de urbanização, quanto ao tamanho dos municípios por faixas populacionais. No entanto, quanto aos serviços de coleta e tratamento de esgotos, a grande concentração espacial da oferta nos aglomerados urbanos do Sudeste e áreas contíguas nos leva a buscar outros fatores que possam ter contribuído para a formação desse padrão.

Estudos recentes sobre a urbanização brasileira podem nos ajudar a explicar o porquê desse padrão. Utilizando o conceito de regiões de articulação ampliada do IBGE, Moura, Oliveira e Pêgo (2018) elaboraram um estudo sobre a urbanização brasileira no qual os municípios foram classificados em quatro escalas de urbanização:

- municípios em estágio mais avançado da urbanização ou em processo de metropolização: cujo conjunto alcançava, em 2010, um grau de urbanização de 93,8%, perfaziam 32% dos municípios brasileiros, abrigavam 72,3% da população e o mais elevado Produto Interno Bruto (PIB) *per capita* entre as escalas: R\$ 24,7 mil;
- municípios fortemente urbanizados: contendo 12,4% do total de municípios brasileiros, a segunda maior proporção de população (19,7%) e do PIB(9,7%);
- municípios urbanizados: o maior número de municípios (42,2% do total), a segunda maior proporção da população (19,7%) e do PIB (9,7%); e

- municípios em transição ou sob influência do urbano: com 13,3% do total de municípios brasileiros, 4,9% do total da população e 1,4% do total do PIB.

Figura 3 – Representação espacial do índice de atendimento urbano por redes de abastecimento de água, segundo faixas percentuais, por município - 2017

Fonte: SANEAMENTO-SNIS (2017b)

Figura 4 – Representação espacial do índice de atendimento urbano por redes coletoras de esgotos, segundo faixas percentuais, por município - 2017

Fonte: SANEAMENTO-SNIS (2017b)

Figura 5 – Taxa de serviços de coleta de resíduos sólidos em relação à população urbana, segundo faixas populacionais - 2017

Fonte: SANEAMENTO-SNIS (2017a)

Na figura 6, que mostra a distribuição espacial dos municípios classificados conforme as escalas de urbanização, podemos verificar que aqueles situados no estágio mais avançado de urbanização formam grandes manchas contínuas no território, particularmente no Sudeste e Sul do País; formam manchas também em continuidade no entorno das metrópoles do Nordeste e Centro-Oeste, ao mesmo tempo em que se espalham em pontos isolados ou manchas menores pelas demais regiões (MOURA; OLIVEIRA; PÊGO, 2018).

De modo geral, observa-se que o Brasil organiza-se a partir de um conjunto de aglomerações fortemente urbanizadas ao longo de todo o contorno litorâneo, adensando-se nas regiões Sudeste e Sul. Nas demais regiões, as centralidades de natureza metropolitana apresentam-se aglomeradas, mas são também evidentes os eixos urbanizados acompanhando o sistema viário principal, no qual se alojam pequenos e médios centros. A imagem de uma rede de cidades isoladas distribuídas no território não encontra correspondência com o urbano brasileiro neste início de século XXI, quando grande parte do cenário, fundamentalmente no contorno leste do País, se compõe por morfologias aglomeradas,

extensas e articuladas espacialmente, que conjugam em grandes manchas contínuas pessoas, atividades e unidades político-administrativas (MOURA; OLIVEIRA; PÊGO, 2018).

Figura 6 – Escalas de urbanização, segundo regiões de articulação ampliada

Fonte: Moura, Oliveira e Pêgo (2018)

2.6 Síntese do problema de pesquisa

Comparando-se as figuras 4 e 6, constata-se notável semelhança entre os padrões espaciais observados, levando-nos a acreditar que haja relação positiva, não verificada nessa pesquisa, entre o fato do município estar situado nas escalas superiores de urbanização e índices mais elevados de atendimento urbano por redes coletoras de esgoto. Mas há outros fatores a serem considerados tais como riqueza e tamanho dos municípios.

Medindo riqueza, em termos de participação no PIB, e o tamanho dos municípios, pela participação no total da população brasileira, a análise de dispersão mostra uma relação positiva entre tamanho do PIB e população. Para melhor visualização, foram excluídos, na figura 7, os municípios com as duas maiores participações no total da população e do PIB:

Sao Paulo e Rio de Janeiro, com respectivamente 11,5% e 5,3% do total do PIB e 5,9% e 3,3% do total da população em 2010. O ponto em destaque corresponde à cidade de Brasília, com 3,7% do PIB e 1,4% da população do Brasil (MOURA; OLIVEIRA; PÊGO, 2018).

Figura 7 – Brasil: participação dos municípios no população total e no PIB - 2010

Fonte: Moura, Oliveira e Pêgo (2018)

Como riqueza e tamanho estão relacionados, é possível focalizar apenas o tamanho dos municípios e sua influência no acesso ao saneamento básico.

Tal como consta no Referencial Teórico, é frequentemente assumido que o tamanho das cidades influencia a provisão de serviços de saneamento básico, especialmente as redes físicas de serviços. De fato, à medida que as cidades crescem em tamanho, devemos esperar ganhos (ou economias) de escala proporcionais ao volume da infraestrutura de saneamento básico.

Usando a população $N(t)$, como medida do tamanho de uma cidade no tempo t , a lei de escala para a infraestrutura pode ser expressa como na equação 1 (BETTENCOURT *et al.*, 2007; BETTENCOURT, 2013):

$$Y(t) = Y_0 N(t)^\beta \quad (1)$$

onde $\beta \approx 0.8 < 1$, Y denota o volume de infraestrutura, e Y_0 é uma constante.

Economias de escala na infraestrutura de saneamento básico implicam redução nos custos, de forma proporcional ao tamanho das cidades, levando a uma (esperada) lei de escala que relaciona gastos com saneamento básico ao tamanho das cidades. Se esta redução nos gastos, proporcional ao tamanho das cidades, não for verificada na prática, significa que os municípios maiores e mais ricos estariam recebendo mais recursos do que o necessário, em detrimento dos municípios menores e mais pobres, o que explicaria a concentração espacial no Sudeste dos serviços de saneamento básico.

A existência da lei de escala nos gastos com saneamento básico é uma hipótese a ser testada por meio de ferramentas de análise estatística aplicadas aos gastos com saneamento urbano. Idealmente deveríamos analisar todos os gastos com saneamento básico, em todos os 5.570 municípios brasileiros, quaisquer que fossem as fontes de recursos, mas, devido à complexidade da tarefa, foi necessário restringir o escopo da pesquisa.

Sendo assim, analisaremos somente os recursos sob gestão da União Federal, onerosos e não onerosos, repassados a fundo perdido aos municípios, ou por meio de operações de crédito aos operadores do sistema. Ou seja, não serão computadas a execução direta, nem as transferências da União Federal aos estados para saneamento básico em municípios, tampouco os recursos próprios dos operadores do sistema. Como uma parte dos recursos declarados como próprios pelos operadores pode ter origem nas transferências de recursos, onerosos e não onerosos, da União Federal, e também dos municípios e estados, compensam-se eventuais perdas decorrentes dessa redução de escopo.

2.7 Objetivos da Pesquisa

2.7.1 Objetivo geral

Explicar a concentração espacial dos serviços de saneamento básico no Brasil, a partir da análise dos padrões de escalamento das transferências, da União Federal para os municípios brasileiros, na modalidade saneamento básico, em relação à população destes municípios.

2.7.2 Objetivos específicos

1. Desenvolver um *software* para extrair dados do DOU, a serem utilizados na detecção de padrões de escalamento nas transferências de recursos, da União Federal para municípios, na modalidade saneamento básico; e
2. Detectar e analisar padrões de escalamento das transferências de recursos, da União Federal para os municípios brasileiros, na modalidade saneamento básico, em relação à população destes municípios.

2.7.3 Questões propostas

- Que dados podem ser usados para detectar padrões de escalamento nas transferências, da União Federal para municípios, na modalidade saneamento básico? Quais as principais características desses dados? Com que frequência são divulgados? São dados georreferenciados?
- Dadas essas características, quais ferramentas de processamento de linguagem natural seriam adequadas à mineração dos dados?
- Dadas essas características, quais algoritmos de aprendizado automático seriam adequados à detecção de padrões nessas transferências?
- Os dados, referentes ao período analisado, mostram quais padrões de escalamento das transferências em função do tamanho dos municípios?
- De que forma os resultados obtidos podem contribuir para o aperfeiçoamento das políticas públicas de saneamento urbano?
- Que outras metodologias podem ser utilizadas para ampliar o controle social e aperfeiçoamento da gestão das políticas de saneamento urbano?

3 Referencial teórico

3.1 Cidades como sistemas complexos

3.1.1 Teoria geral de sistemas e cibernética

Nesse item vamos rever alguns aspectos da teoria geral de sistemas e da cibernética que passaram a ser aplicados ao planejamento e gestão de cidades, a partir das décadas de 1950 e 1960, bem como as mudanças recentes introduzidas pelas ciências da complexidade.

Cidades começaram a ser vistas como sistemas quando a teoria geral de sistemas e a cibernética foram aplicadas às ciências sociais, há mais de meio século. Bertalanffy (1968), na biologia, e Wiener (1948), na engenharia, deram enorme impulso a esta área interdisciplinar, e disseminaram a ideia de que fenômenos de interesse de várias disciplinas poderiam ser articulados como "sistemas" e que procedimentos de planejamento, controle e gestão de sistemas poderiam ser aplicados às mais diversas áreas do conhecimento. Da teoria geral de sistemas provinha a lógica geral para explicar a estrutura e o comportamento dos sistemas, por meio de várias formas de retroalimentação e organização hierárquica, enquanto a cibernética era "a ciência da navegação", que possibilitava a esses sistemas moverem-se em direção a objetivos ou metas explícitas (BATTY, 2008).

Cidades correspondiam perfeitamente a essa caracterização e, nas décadas de 1950 e 1960, a abordagem tradicional, segundo a qual cidades seriam estruturas articuladas que requeriam organização estética e física, rapidamente deu lugar à abordagem de cidades como sistemas complexos. Sendo assim, seus controle e planejamento requeriam intervenções bem mais sutis que qualquer coisa que tivesse sido feita até então na área de planejamento urbano (BATTY, 2008).

Tornava-se claro que sistemas sociais complexos - corporações, cidades, economias, e governos - seriam contra-intuitivos. Ações corretivas aplicadas a estes sistemas poderiam se mostrar frequentemente ineficazes, e seus resultados adversos. Com frequência, políticas adotadas para sanar determinadas dificuldades acabavam por intensificá-las. A explicação é que, em sistemas complexos, causa e efeito quase nunca estão relacionados no tempo e no espaço (FORRESTER, 1970).

Na teoria geral de sistemas, a estrutura de sistemas complexos, tais como cidades, não é constituída por um laço simples de retroalimentação (*feedback loop*), no qual um único estado domina o comportamento do sistema (figura 8).

Figura 8 – Laço de retroalimentação o mais simples possível com uma taxa e um nível

Fonte: Forrester (1970)

Sistemas complexos possuem uma multiplicidade de laços de retroalimentação que interagem. Suas taxas de fluxo são controladas por relações não-lineares. Sistemas complexos possuem muitos estados ou níveis. Eles possuem laços de retroalimentação positivos que geram crescimento, amplificam desvios e reforçam mudanças. E também laços de retroalimentação negativos que agem para conduzir o estado do sistema ao estado ou objetivo desejado. Todos os laços de retroalimentação negativos têm a estrutura mostrada na figura 9.

Figura 9 – Estrutura dos laços de retroalimentação negativos

Fonte: Sterman (2000)

Em sistemas complexos, a causa de uma dificuldade pode estar localizada atrás dos seus sintomas, no tempo, ou em uma parte do sistema completamente diferente e remota, no espaço. De fato, causas são usualmente encontradas, não nos eventos precedentes, mas sim na estrutura e políticas do sistema. Condicionados por nossa experiência na análise de sistemas simples, aplicamos a mesma intuição aos sistemas complexos, o que nos leva a cometer erros. Os resultados são inefetivos ou até mesmo prejudiciais (FORRESTER, 1970).

Forrester (1970), e outros autores que usaram a teoria geral de sistemas para explicar a dinâmica urbana, afirmavam que cidades podiam ser representadas como conjuntos de elementos ligados por uma série de interações. Para estes autores, a estrutura das cidades baseava-se na forma das atividades de uso da terra e nas ligações econômicas e funcionais entre elas, representadas inicialmente em termos de movimentos físicos, e de tráfego. A dinâmica de retroalimentação, que mantém o sistema integrado, era representada em termos de volume e padrão destas interações, em determinado ponto no tempo. A evolução de longo prazo da estrutura urbana não era central nessa abordagem, porque seu foco principal era o funcionamento das cidades como estruturas em equilíbrio. Seu objetivo era melhorar as interações entre os usos da terra, tornando-as mais eficientes, e ao mesmo tempo buscar a equidade social e espacial. Transporte e habitação eram de importância central na defesa do argumento de que cidades deveriam ser tratadas como exemplos de sistemas gerais e ajustadas de acordo com os princípios da cibernética (BATTY, 2008).

Embora os modelos baseados nessa premissa ainda sejam úteis, a tendência atual é não mais tratar as cidades como estruturas em equilíbrio. A percepção de que as cidades sejam mais propensas ao desequilíbrio, ou que devam ser classificadas como sistemas longe do equilíbrio, é coerente com a velocidade da mudança e a volatilidade observadas nos últimos cinquenta anos. Estruturas urbanas em equilíbrio estão sendo rompidas por inovações inesperadas, muitas de ordem tecnológica, outras de ordem social, que alteram a forma como as pessoas tomam decisões sobre onde morar e como se mover nas cidades. A ideia de que estas novas estruturas e comportamentos emergentes sejam com frequência inesperadas integra o paradigma emergente. Cidades planejadas são sempre a exceção e não a regra, e quando são diretamente planejadas, só se mantêm assim por períodos muito curtos de tempo (BATTY, 2008).

As novas ciências da complexidade estão reformulando a teoria geral de sistemas, embora ainda estejam fundamentadas em estruturas compostas por elementos, denominados

atores ou agentes, cujas interações determinam os processos comportamentais que mantêm o sistema em equilíbrio ou o deslocam para novos estados. Nesta nova abordagem, os mecanismos de retroalimentação ainda são centrais, mas o objetivo agora é entender como os elementos reagem uns com os outros através do tempo. O conceito de equilíbrio estável não é mais considerado central; a retroalimentação é vista como a forma pela qual estas estruturas evoluem para novos estados. Ou seja, a retroalimentação passou a ser vista no sentido positivo, em vez de negativo, embora ambos sejam essenciais. As interações entre os elementos do sistema foram enriquecidas por novos conceitos de redes e sua dinâmica (Newman, Barabasi e Watts (2011)). Noções sobre como os elementos escalonam uns em relação aos outros e em relação aos níveis hierárquicos do sistema se tornaram úteis para mostrar como as interações locais levam a padrões globais que somente podem ser previstos de baixo para cima (Miller e Page (2007)). Estes padrões emergentes podem ser reproduzidos por modelos que mostram o crescimento das cidades de baixo para cima (BATTY, 2008).

Ao analisarmos as morfologias urbanas, à luz dessa nova abordagem, elas nos parecem caóticas, mas ordenadas, auto-similares através de múltiplas escalas, mas crescendo organicamente de baixo para cima. Sobre a auto-similaridade em múltiplas escalas, as evidências obtidas a partir dos estudos empíricos realizados nos últimos quarenta anos mostram que a maioria das propriedades urbanas, denotadas por Y , variam continuamente com o tamanho da população e são bem descritas, na média, por leis de potência da forma $Y = Y_0 * N^\beta$; onde Y_0 e β são constantes em N . O mais surpreendente é que cidades com diferentes tamanhos de população tenham de fato propriedades diferentes (BETTENCOURT, 2013).

Especificamente, observa-se que taxas de quantidades sociais (produtos, salários ou novas invenções) aumentam, *per capita*, com o tamanho da cidade (escalamento super-linear com $\beta = 1 + \delta > 1$ e $\delta \approx 0.15$), enquanto o volume ocupado pela infraestrutura *per capita* (estradas, cabos, etc) decresce (escalamento sublinear com $\beta = 1 - \delta < 1$). Estes resultados corroboram a impressão de que cidades maiores não são apenas mais caras e congestionadas, mas também mais excitantes e criativas, quando comparadas às cidades menores. Estes resultados empíricos também sugerem que, a despeito de sua aparente complexidade, as propriedades médias globais das cidades podem ser definidas por um pequeno conjunto de parâmetros.

A figura 10 mostra dois exemplos de escalamento em cidades. O gráfico da direita mostra o escalamento superlinear do Produto Metropolitano Bruto, em relação ao tamanho da população urbana. Os dados referem-se a 415 regiões metropolitanas dos Estados Unidos no ano de 2006 (pontos verdes). O gráfico da esquerda mostra o escalamento sublinear do volume total em milhas (*lane-miles*) das estradas das regiões metropolitanas dos Estados Unidos (pontos azuis) no mesmo ano (BETTENCOURT, 2013).

Figura 10 – Escalamento sublinear e superlinear em cidades

Fonte: Bettencourt (2013)

Em síntese, o estudo das cidades está sendo transformado por novas abordagens originárias das ciências da complexidade. A teoria de sistemas, que tratava as cidades como se fossem organizadas de cima para baixo, foi substituída por abordagens que mostram sua evolução de baixo para cima. Nesta nova visão, cidades são organismos, não máquinas. Na dinâmica das cidades, a noção de equilíbrio também está sendo substituída por uma mistura de diferentes tipos de caos, catástrofes e bifurcações. A partir das redes e interações que sustentam as cidades é possível mostrar como o movimento e a mobilidade dão origem à difusão e segregação de diferentes atividades espaciais. Ao analisarmos os padrões fractais e de auto-similaridade das morfologias urbanas, identificamos leis de escala baseadas em ordem de frequência, alometria e gravitação. Tais ideias estão sendo incorporadas por modelos que podem subsidiar políticas urbanas (BATTY, 2012).

Vemos agora a emergência de uma nova ciência das cidades. Esta ciência em desenvolvimento não abandonou as abordagens mais formais da economia urbana, da ciência regional, da física social e do planejamento do transporte. Na verdade, ela está

sendo construída sobre o edifício das abordagens mais formais, enquanto altera sua ênfase (BATTY, 2012).

3.1.2 A nova ciência das cidades

Nesse item vamos rever alguns aspectos da nova ciência das cidades, com destaque para as leis de escala e para os modelos de crescimento de redes capazes de explicar as morfologias urbanas.

O funcionamento das cidades no espaço e no tempo é baseado em múltiplos processos de escolha espacial, em que indivíduos e grupos populacionais se localizam uns em relação aos outros e em relação às suas atividades na forma de tipos de uso da terra. Estas atividades tendem a ser dominadas por *trade-offs* entre economias e deseconomias de aglomeração, frequentemente representadas em termos de acessibilidade relativa entre diferentes localidades. Estes *trade-offs* dão origem a padrões de atividade que refletem diferentes níveis de clusterização que, por sua vez, implicam diferentes níveis de densidade associados a diferentes localidades (BATTY, 2013).

Por um longo período, os padrões espaciais das cidades foram representados, em termos grosseiros, como perfis de densidade em torno de *hubs* importantes, hierarquias aninhadas de locais centrais (onde o nível seguinte é composto por duas ou mais unidades do nível anterior), e padrões de acessibilidade refletindo formas policêntricas, mas fracamente relacionados com as estruturas morfológicas reais que deveriam representar. O que significa que os pesquisadores urbanos não percebiam características importantes tais como a auto-similaridade ou a invariância espacial através de diversas escalas, uma evidência de que processos similares estavam operando através de diversas escalas. No entanto, padrões que se repetem de forma modular, tais como aqueles característicos da teoria das localidades centrais (Christaller (1966), Losch *et al.* (1954)), geram hierarquias que podem ser analisadas usando novos tipos de geometria (figura 11). E de fato, na década de 1980, surgiram novas ideias sobre como os padrões modulares eram estruturados e, usando conceitos sobre processos auto-similares de desenvolvimento, cidades começaram a ser vistas como fractais (BATTY; LONGLEY, 1994; BATTY, 2013).

Figura 11 – Sistemas de localidades centrais de Christaller

Fonte: White, Engelen e Uljee (2015)

O desenvolvimento urbano raramente preenche todo o espaço que define uma cidade, e, neste sentido, é análogo à forma como os fractais preenchem o espaço entre as dimensões inteiras. Na realidade, cidades foram reconhecidas como formas fractais por arquitetos, planejadores e outros, antes do conceito de fractal ser formalmente formulado. No entanto, a adaptação do conceito matemático a aplicações do mundo real não é exata nem desprovida de complicações. Por exemplo, uma fronteira fractal gerada matematicamente tem uma dimensão bem-definida $1 < D_b < 2$. Mas se estimássemos o valor de D_b medindo o real comprimento da fronteira teríamos que descer a resoluções cada vez mais finas para extrair a relação entre a resolução do procedimento de medida e o comprimento medido (WHITE; ENGELLEN; ULJEE, 2015).

Cidades são fractais sob diversos aspectos. Os limites da área construída, a distribuição espacial dos usos da terra, a distribuição por tamanho dos *clusters* internos, as redes de transportes, e várias características locais são fractais. Dado que as qualidades fractais das cidades representam características gerais da estrutura urbana, as dimensões fractais, que determinam a extensão em que as cidades preenchem o espaço que ocupam, são importantes tanto para a compreensão teórica dos processos urbanos quanto para calibrar e validar os modelos desses processos; podem também ser importantes no planejamento e formulação de políticas urbanas (WHITE; ENGELEN; ULJEE, 2015).

A **dimensão radial** descreve a relação entre a área construída e a distância (raio), calculada a partir do centro da cidade. Se a área inteira for construída, então ela crescerá com o quadrado da distância a partir do centro, mas como, em geral, a área construída se torna menos densa à medida que a distância a partir do centro cresce, ela crescerá somente com uma potência $D_r < 2$ da distância a partir do centro, onde D_r é a dimensão radial. No caso mais simples, supondo uma cidade que cresça a partir do centro, essa relação pode ser escrita como na equação 2 (WHITE; ENGELEN; ULJEE, 2015):

$$A(\varepsilon) = k\varepsilon^D \quad (2)$$

Se calcularmos o logaritmo dos dois lados da equação, obteremos $\log(A(\varepsilon)) = D\log(\varepsilon)$ onde $A(\varepsilon)$ é a área de um objeto urbano dentro da distância ε a partir do centro, e o expoente D é a dimensão radial. O coeficiente k foi incluído para abarcar diversos fenômenos. Se $D < 2$, o objeto urbano ocupará cada vez menos da terra disponível à medida que a distância do centro crescer. Se, contudo, $D > 2$, o uso da terra ocupará progressivamente mais terra disponível à medida que a distância do centro aumentar; neste caso, $k < 1$ (WHITE; ENGELEN; ULJEE, 2015).

Por sua vez, a **dimensão de correlação** é uma generalização da dimensão radial, na qual cada célula de um objeto urbano é considerada um centro, a partir do qual a distância ε é aumentada incrementalmente. Se um objeto urbano tiver N células, para cada incremento de ε , existem N valores de $A(\varepsilon)$, para os quais podemos calcular o valor médio $\langle A(\varepsilon) \rangle$ (WHITE; ENGELEN; ULJEE, 2015).

Finalmente, a **dimensão de tamanho-frequência dos clusters**, também conhecida como "ruído $1/f$ " descreve a relação entre o tamanho de um *cluster* e o número de *clusters* daquele mesmo tamanho encontrados no sistema. Ela é usada para caracterizar a

distribuição por tamanho dos clusters urbanos em uma região ou de *clusters* de um tipo particular de uso da terra em cidades. Essa relação pode ser escrita como na equação 3:

$$N(s) = ks^{-D} \quad (3)$$

Se calcularmos o logaritmo dos dois lados da equação obteremos uma relação linear da forma:

$$\log N(s) = \log(k) - D \log(s) \quad (4)$$

onde $N(s)$ é o número de clusters por tamanho, e D é a dimensão de tamanho-frequência dos clusters tal como na figura 12. Se a relação não for linear, então o sistema não é fractal sob esse aspecto.

Figura 12 – Gráfico tamanho-frequência dos *clusters* para a Holanda - 1994

Fonte: :
White, Engelen e Uljee (2015)

As dimensões fractais são um meio de classificar as cidades segundo suas densidades relativas. Muitas destas ideias podem ser expressas em termos de estruturas de rede que interligam os diversos usos da terra e geram o tipo de economias e efeitos de escala que definem as cidades (BATTY; LONGLEY, 1994; BATTY, 2013).

Os vários processos que ocorrem no meio urbano definem múltiplas redes que permitem que as populações entreguem materiais e informação para suas atividades. Redes, físicas e sociais, tendem a ser reforçar mutuamente e assim, a partir de um *hub* inicial, indivíduos são atraídos de forma proporcional à estrutura existente.

Esse mecanismo de crescimento de redes, denominado ligação preferencial, foi proposto por Barabási e Albert (1999), como parte de seu modelo de redes a ser detalhado no próximo item. Uma de suas principais características é que a distribuição dos vértices por tamanho segue uma lei de escala (ou de potência) do tipo:

$$Prob(X = x) \simeq x^{-k} \quad (5)$$

Essa lei de escala é aplicável não só a redes, mas também aos tamanhos das cidades, e ainda aos tamanhos de diferentes localidades dentro das cidades. A distribuição de frequência das cidades é fortemente enviesada para a esquerda, podendo ser aproximada por uma lei de potência inversa, que, por sua vez, equivale à Lei de Zipf. A Lei de Zipf estabelece que a frequência X de ocorrência de um evento é inversamente proporcional à sua classificação por tamanho r (GABAIX, 1999).

3.1.3 Leis de escala em cidades

Batty (2013) enumerou sete leis de escala aplicáveis às cidades:

1. À medida que as cidades crescem, o número de conexões potenciais cresce aproximadamente com o quadrado da população. Na computação, esta relação é conhecida como **Lei de Metcalfe** que trata do valor dos sistemas de comunicação em termos do número de usuários ou nós do sistema n e pode ser expressa matematicamente como:

$$Valor = \frac{n(n-1)}{2} \approx n^2 \quad (6)$$

2. À medida que as cidades tornam-se maiores, sua renda média real aumenta mais que proporcionalmente com a população, com não-linearidade positiva. Esta relação pode ser denominada Lei de Bettencourt-West, embora também possa ser denominada Lei de Marshall, autor do conceito de economias de aglomeração em cidades, no século XIX. Como visto anteriormente, a **Lei de Bettencourt-West** pode ser expressa da forma:

$$Y = Y_0 N^\beta \quad (7)$$

3. À medida que as cidades tornam-se maiores, há menos exemplares delas. Esta relação corresponde à **Lei de Zipf**:

$$X \simeq r^{-b} \quad (8)$$

onde X é a frequência da ocorrência de um evento, r corresponde à classificação por tamanho e b é um valor muito próximo de 1.

4. À medida que as cidades crescem, em torno do seu ponto central, que pode ser a praça do mercado ou a sede do governo, várias densidades, mais especificamente o preço da terra e a densidade de ocupação, declinam de forma não-linear com a distância ou custo de deslocamento a partir de seu ponto central. Esta relação pode ser denominada **Lei de von Thunen**. Sua inversa é a relação massa-raio, uma relação fractal, que estabelece que a massa da população cresce de forma superlinear com o crescimento da distância, a partir do distrito central de negócios, com um expoente(dimensão fractal) entre 1 e 2 (BATTY; LONGLEY, 1994).
5. À medida que as cidades tornam-se maiores, as interações entre elas escalam com algum produto do seu tamanho, mas as interações declinam com o crescimento da distância ou custo do deslocamento. Esta relação pode ser considerada uma versão bidimensional da Lei de von Thunen. Mas, na verdade, ela foi enunciada antes por Tobler, na forma "Tudo está relacionado com tudo, mas as coisas mais próximas estão mais relacionadas que as distantes", sendo conhecida como **Primeira Lei da Geografia**(TOBLER; MIELKE; DETWYLER, 1970; SUI, 2004).
6. À medida que as cidades tornam-se maiores, suas áreas centrais declinam em termos de densidade populacional e seus perfis de densidade se achatam. Isto tem sido observado em muitas cidades, em especial na cidade de Paris, por Bussière, devendo ser denominada **Lei de Bussière**.
7. À medida que as cidades tornam-se maiores, elas também se tornam mais sustentáveis. Esta relação foi enunciada por Brand (2010) cujas observações sugerem ser necessário que as cidades tornem-se mais densas para uma boa interação social e para tornar mais custo-efetivas e sustentáveis as soluções de transporte e de construção. A **Lei de Brand** é uma descoberta recente, sendo de fato contraditória com a dinâmica das cidades na segunda Revolução Industrial e de algumas cidades do mundo em desenvolvimento.

Leis de escala são úteis para detectar os processos subjacentes à geração e ao crescimento de estruturas urbanas. Padrões similares, fractais e livres de escala, têm sido observados em muitas redes sociais humanas, não só em cidades. Portanto, é natural que as características essenciais de uma teoria quantitativa e preditiva das cidades se

originem na dinâmica das redes de infraestrutura e sociais, e que estas sustentem o escalamento observado e os valores universais aproximados dos expoentes das leis de escala (BETTENCOURT *et al.*, 2010).

3.1.4 Modelos de redes aplicados a cidades

Neste item faremos inicialmente uma revisão dos principais modelos de crescimento de redes, para em seguida mostrarmos como o mecanismo de ligação preferencial, do modelo de Barabási-Albert, pode explicar o surgimento de algumas morfologias urbanas.

Uma rede é um conjunto de vértices ou nós, com conexões entre eles, denominadas arestas ou ligações. Redes são abundantes no mundo: a internet, redes sociais, redes organizacionais e redes de relações de negócios entre empresas, redes neurais, redes metabólicas, redes de alimentos, redes de distribuição, tais como vasos sanguíneos ou rotas de serviços postais, redes de citações de artigos, e muitas outras (NEWMAN, 2003).

De forma geral, podemos dizer que uma rede ou grafo $G(n, g)$ é composto por um grupo de vértices ou nós $N = \{1, \dots, n\}$ e uma matriz de valores reais g de dimensão $n \times n$, onde g_{ij} representa a relação (direta/indireta) entre i e j . Essa matriz é conhecida como matriz de adjacências (*adjacency matrix*), porque ela lista quais vértices são adjacentes um ao outro. No caso das entradas de g poderem usar mais do que dois valores e indicar a intensidade dos relacionamentos, o grafo é denominado grafo ponderado (*weighted graph*). Mas se, ao contrário, os valores de g forem 0 ou 1, então o grafo é não-ponderado (*unweighted graph*) (JACKSON, 2010).

Uma rede é direta se for possível que $g_{ij} \neq g_{ji}$ e indireta se $g_{ij} = g_{ji}$ para todos os vértices i e j . Por exemplo, se $N = 1, 2, 3$, então:

$$g = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad (9)$$

É uma rede, indireta e não-ponderada, com uma aresta entre os vértices 1 e 2, outra aresta entre 2 e 3, mas nenhuma entre os vértices 1 e 3 (JACKSON, 2010).

Redes Aleatórias: Modelo de Erdős-Rényi

Erdos e Rényi (1960) propuseram um dos primeiros modelos teóricos de crescimento de redes: a rede ou grafo aleatório. A versão mais conhecida desse modelo, G_{np} , é tal que cada aresta possível entre dois vértices está presente com probabilidade p , e ausente com probabilidade $(1 - p)$. A probabilidade de que um nó tenha exatamente grau k é dada por uma distribuição binomial. Para um número grande de nós, $n \gg kz$, onde z é o grau médio da rede, ela pode ser aproximada por uma distribuição de Poisson (NEWMAN, 2002).:

$$p_k = \frac{z^k e^{-z}}{k!} \quad (10)$$

Quando $p > 1/n$, onde n é o número de vértices, então é possível observar a emergência de um componente gigante. Um componente gigante é um subconjunto de vértices do grafo alcançável a partir dos outros vértices, por meio de algum caminho na rede.

O grafo aleatório de Erdős e Rényi possui a vantagem de fornecer a solução exata para várias de suas propriedades médias. No entanto, ele deixa a desejar como representação de redes do mundo real, pois falha em duas questões cruciais: (i) clusterização (*clustering*) incompatível com essa mesma característica das redes do mundo real; e (ii) distribuição de graus de Poisson irreal (NEWMAN, 2002).

O coeficiente de clusterização C pode ser definido como a probabilidade média de que dois vizinhos de um dado vértice também sejam vizinhos um do outro. Nas redes do mundo real, o coeficiente de clusterização medido tem um valor alto, próximo de 50% por cento ou ainda mais. No entanto, no modelo teórico de Erdős-Rényi, as probabilidades de pares de vértices estarem conectados por arestas são, por definição, independentes, de tal modo que não é mais provável que estejam conectados se tiverem um vizinho comum, mas sim o contrário. Isso significa que o coeficiente de clusterização de um grafo aleatório é dado simplesmente por $C = p$ ou, aproximadamente, $C = z/n$, onde z é o grau médio da rede (NEWMAN, 2002).

Além da discrepância entre os coeficientes de clusterização, das redes aleatórias quando comparadas às reais, constata-se também incongruência entre a distribuição de graus das redes reais e a distribuição de Poisson que subestima significativamente a quantidade de nós com alto grau. Por exemplo, o grau máximo esperado da internet seria

mais ou menos 20. Mas os dados empíricos indicam a existência de nós com graus próximos a 10^3 . Por outro lado, a dispersão nos graus das redes reais é muito maior que a esperada em uma rede aleatória. Se a internet fosse uma rede aleatória, esperaríamos uma dispersão da ordem de 2.52. No entanto, as medições apontam uma dispersão da internet bem maior, da ordem de 14.14 (BARABÁSI, 2014).

Redes de Pequeno Mundo: Modelo de Watts-Strogatz

Vimos que o coeficiente de clusterização das redes do mundo real está próximo de 50%. Outra característica destas redes são caminhos médios mais curtos, fenômeno conhecido como "mundo pequeno", ou ainda "seis graus de separação". Por "seis graus de separação", queremos dizer que, para quaisquer dois indivíduos no planeta Terra, é possível encontrar um caminho entre eles com, no máximo, seis conhecidos (GUARE, 1990).

Na linguagem da ciência das redes, o fenômeno "mundo pequeno" implica que a distância entre dois nós escolhidos aleatoriamente é curta. Mas curta comparada a que? Para tornar este conceito mais preciso, definimos que a distância máxima ou diâmetro de uma rede aleatória d_{max} pode ser aproximada por:

$$d_{max} \approx \frac{\ln N}{\ln \langle k \rangle} \quad (11)$$

onde N é o número de nós e $\langle k \rangle$ é o grau médio (denominado anteriormente z).

Essa equação prediz o escalamento do diâmetro da rede com o tamanho N do sistema. Para ilustrar suas implicações em redes reais, vamos usar $N = 7 \times 10^9$ (número de indivíduos no planeta Terra) e $\langle k \rangle = 10^3$ o que nos permite obter $\langle d \rangle \approx 3.28$. Portanto, todos os indivíduos no planeta Terra devem estar situados dentro do intervalo de 3 a 4 apertos de mãos uns dos outros. Provavelmente essa é uma estimativa mais realista que os famosos 6 graus de separação (BARABÁSI, 2014):

O modelo de Watts-Strogatz interpola entre uma rede regular, que possui alta clusterização, mas não consegue reproduzir o fenômeno do mundo pequeno, e uma rede aleatória, que tem baixa clusterização, mas reproduzi-lo. Por ser uma extensão do modelo de redes aleatórias, o modelo de Watts-Strogatz prediz uma distribuição de graus parecida com a distribuição de Poisson. Consequentemente, nós com graus mais altos não estão corretamente representados nesse modelo (BARABÁSI, 2014).

Redes Invariantes de Escala: Modelo de Barabási-Albert

Se a internet fosse uma rede aleatória, os graus dos seus vértices seguiriam uma distribuição de Poisson. No entanto, vemos que se os dados da distribuição de graus da internet fossem representados na escala log-log, eles formariam uma linha aproximadamente reta, sugerindo que a distribuição pode ser aproximada por:

$$p_k \sim k^{-\gamma} \quad (12)$$

Essa equação é uma distribuição de lei de potência e γ é o seu expoente de grau. Se calcularmos o logaritmo de ambos os lados da equação obteremos: $\log(p_k) \sim -\gamma \log(k)$. Logo espera-se que p_k dependa linearmente do $\log(k)$. A inclinação desta reta é o expoente de grau γ . Ou seja, os resultados empíricos mostram uma rede cuja distribuição de graus é completamente diferente da distribuição de Poisson. Essas redes são denominadas livres de escala. Uma rede livre de escala é uma rede cuja distribuição de graus segue uma lei de potência (BARABÁSI, 2014).

A rede se expande continuamente por meio da adição de novos vértices que se conectam aos existentes no sistema. Mas, diferentemente do que ocorre na rede de Erdős-Rényi, a probabilidade de que dois vértices se conectem não é aleatória nem uniforme, mas sim proporcional ao grau do vértice já existente. Esse mecanismo é denominado ligação preferencial (*preferential attachment*). Mais precisamente, a probabilidade de que um novo vértice se conecte ao vértice i depende do grau do vértice i e é dada por:

$$P(k_i) = \frac{k_i}{\sum_j k_j} \quad (13)$$

Por sua vez, a distribuição de graus do modelo de Barabási-Albert obedece a uma lei de escala do tipo:

$$p(k) \sim 2m^{1/\beta} k^{-\gamma} \quad (14)$$

onde γ independe de m , número de nós já existentes quando um novo nó se conecta à rede; e k = taxa em que cada nó i adquire uma ligação. As Figuras 13 e 14 mostram uma rede gerada com $m = 120$ nós, $\gamma = 3$ e $k = 4$ e sua respectiva distribuição de graus.

Figura 13 – Modelo de Barabási-Albert: grafo

Fonte: Ludmila Deute Ribeiro, 2018

Figura 14 – Modelo de Barabási-Albert: histograma

Fonte: Ludmila Deute Ribeiro, 2018

Neste tipo de rede, $p(k)$ independe do tempo e consequentemente do tamanho do sistema $m_0 + t$ onde m_0 é o número inicial de nós ou vértices. Ou seja, apesar do crescimento contínuo, o sistema se organiza em um estado independente de escala. As redes livres de escala possuem clusterização mais alta e caminhos mais curtos entre os nós que as redes aleatórias.

Ligação Preferencial e Lei de Zipf em Cidades

Os processos que levam ao crescimento de redes também podem explicar as hierarquias de tamanho entre cidades e localidades dentro de cidades, tal como expresso na Lei de Zipf.

Supondo que subsistemas urbanos (localidades) possam ser representados por vértices, e os fluxos entre eles por arestas, o crescimento de cidades pode ser simulado, por exemplo, por um modelo desenvolvido por Batty (2013), descrito a seguir.

A cada período de tempo, para um vértice já estabelecido e ligado a outros vértices, vamos considerar a adição aleatória, de um volume de arestas dado por $\delta_{ij}(t+1) = 1$ a $P_{ij}(t)$, ou seja, ao número total de arestas ligando o vértice i ao vértice j no tempo t . Dessa forma, no tempo $t+1$, o número total de arestas associadas a i será dado por:

$$P_i(t+1) = \sum_k P_{ik}(t) + \delta_{ij}(t+1) \quad (15)$$

onde o número de arestas é atualizado a cada período como:

$$P_{ij}(t+1) = P_{ij}(t) + \delta_{ij}(t+1) \quad (16)$$

Se uma aresta será adicionada ou não depende do tamanho do vértice e de sua distância até outros vértices, o que é refletido por uma função gravitacional exponencialmente ponderada da forma:

$$\delta_i(t+1) = \begin{cases} 1, & \text{se } rnd(\zeta_{ij}(t+1)) > KP_{ij}exp(-\varphi_{ij}) \\ 0 & \end{cases} \quad (17)$$

O termo $rnd(\zeta_{ij}(t+1))$ determina a escolha aleatória com base na intensidade da interação potencial, onde d_{ij} é a distância do vértice i para o vértice j e o parâmetro φ reflete o efeito friccional da distância. Essencialmente, este processo corresponde à ligação preferencial no qual arestas são adicionadas de forma proporcional à quantidade de arestas existentes e à população que o vértice já tiver atraído.

O processo de ligação preferencial foi usado por Albert e Barabási (2002) para mostrar que o modelo gera redes livres de escala nas quais o número de graus de ligações de entrada e de ligações de saída escalam de acordo com uma lei de potência.

Ao processo de ligação preferencial Batty (2013) acrescentou um mecanismo para explicar o nascimento de novos vértices. Podemos estabelecer que um novo vértice em

i será adicionado se a variável aleatória $rnd(v(t+1))$ for maior que um determinado patamar z , conforme equação 18:

$$\delta_i(t+1) = \begin{cases} 1, & \text{se } rnd(v_{ij}(t+1)) > z \\ 0 & \end{cases} \quad (18)$$

onde o valor de z é pequeno comparado à probabilidade de novas arestas.

Para executar o modelo, Batty (2013) usou uma grade de 21×21 e um patamar $z = 0.1$, o que significa que, no início de cada período de tempo, há 1 chance em 10 de um novo vértice ser adicionado. Se o processo continuar, esta chance diminuirá, pois se o vértice escolhido já estiver estabelecido, ele será abandonado. Para gerar arestas entre os vértices estabelecidos, um vértice é primeiro escolhido aleatoriamente, mas em proporção ao seu tamanho $P_i(t)$, e uma aresta para outro vértice j a partir de i é escolhida proporcionalmente à função inversa da distância tal como na equação 18. Desta forma, a rede vai crescendo por meio da ligação preferencial dos novos vértices àqueles já estabelecidos.

A figura 15 ilustra a distribuição final da população por vértice $P_i(1000)$ e também a distribuição dos volumes de arestas entre todos os vértices maiores que 1, maiores que 2, e finalmente os maiores que 4. Estas figuras mostram um padrão hierárquico entre diferentes áreas, de acordo com sua densidade de clusterização. A figura 16 é um gráfico da população por vértice de acordo com sua classificação. O gráfico é aproximadamente lognormal, semelhante à lei de potência. Se ajustarmos uma reta, obteremos um expoente de 1.05 muito próximo ao da Lei de Zipf onde b é aproximadamente 1 (BATTY, 2013).

Batty (2013) sustenta que muitos sistemas reais possuem parâmetros semelhantes aos do modelo descrito. Um exemplo seria a hierarquia das atividades de varejo na cidade de Londres, representada por um índice de intensidade de varejo calculado como uma soma linear ponderada de 42 indicadores separados, cada um deles devidamente normalizado, e georreferenciado, e que, na sua escala mais fina, mostra as atividades de varejo com resolução espacial de cerca de 50 metros. Interpolando uma superfície a partir desses dados e recortando-a em 5 diferentes níveis, obteve-se uma imagem da hierarquia do varejo mostrada na Figura 17 (BATTY *et al.*, 2002).

Figura 15 – Padrões de conectividade de rede

Fonte: Batty (2013)

Figura 16 – Distribuição dos nós da rede por classificação

Fonte: Batty (2013)

Cidades são sistemas complexos por excelência, representam mais que a soma de suas partes, e são desenvolvidas por meio de uma multiplicidade de decisões coletivas e individuais de baixo para cima e de cima para baixo. As novas ciências da complexidade são importantes para o seu entendimento. No entanto, este entendimento torna-se um objetivo cada vez mais desafiador e móvel porque as próprias cidades estão se tornando mais complexas devido às tecnologias que estamos utilizando para compreendê-las (BATTY *et al.*, 2012).

O conceito de cidades inteligentes emergiu durante a última década como uma fusão de ideias sobre como as tecnologias de informação e comunicação (TIC) poderiam melhorar o funcionamento de cidades, aumentando sua eficiência e competitividade, e gerando novas formas de resolver os problemas de pobreza, privações sociais e más condições ambientais. O termo "cidades inteligentes" tem muitas faces. Cidades inteligentes, cidades virtuais, cidades digitais, cidades da informação são diferentes perspectivas da ideia de que TICs são centrais para a operação da cidade do futuro (BATTY *et al.*, 2012).

No próximo item vamos analisar o conceito de cidades inteligentes e sua conexão com o big data, ou seja, grandes volumes de dados gerados de forma automática por meio do uso de tecnologias de informação e comunicação.

Figura 17 – A hierarquia implícita do varejo no centro de Londres

Fonte: Batty (2013)

3.2 *Big data e planejamento urbano*

Nesse item vamos tentar definir o que são cidades inteligentes, *big data*) e as relações entre ambos os conceitos. Nosso objetivo é avaliar possibilidades e limitações de uso de *big data* no planejamento urbano.

Atualmente muitos governos estão planejando desenvolver cidades inteligentes para garantir a necessária sustentabilidade e elevar o padrão dos serviços urbanos. Mas afinal o que são cidades inteligentes? Ainda não existe uma definição precisa do termo, no entanto, a partir dos projetos existentes, podemos definir cidades inteligentes como uma solução integrada de moradia que conjuga vários aspectos tais como energia, transporte e edificações, de uma forma inteligente e eficiente, visando a melhoria da qualidade de vida dos cidadãos. Além disso, os projetos também enfatizam a importância da sustentabilidade de recursos e aplicações para as futuras gerações. É possível observar estes aspectos em cada proposta de cidade inteligente, a despeito do tamanho, localização ou recursos disponíveis (NUAIMI *et al.*, 2015).

O desenvolvimento das cidades inteligentes baseia-se na utilização de *hardware* e software de tecnologias de informação, redes e dados sobre diferentes áreas e serviços. Também abrange componentes tais como recursos naturais, infraestrutura, energia, transporte, educação, cuidados com a saúde, governo e segurança pública. A disponibilidade de recursos, financeiros, naturais ou humanos, é um dos maiores desafios à construção e manutenção de uma cidade inteligente. Sistemas regulatórios também podem afetar suas chances de sucesso. Acima de tudo, existem os desafios tecnológicos que requerem soluções avançadas. Por outro lado, novas e emergentes tecnologias podem ajudar a transformar desafios em oportunidades (NUAIMI *et al.*, 2015).

Big data é uma tecnologia emergente que pode ser usada para esse fim. À medida que a digitalização torna-se parte integrante da vida cotidiana, há um acúmulo de grandes volumes de dados que podem ser usados em várias aplicações urbanas. Os dados são gerados por múltiplas fontes: telefones inteligentes, computadores, sensores ambientais, câmeras, Sistema de Geoposicionamento Global (GPS), ou mesmo pessoas. Mídias sociais, imagens digitais e vídeos, transações comerciais, jogos, dentre outros, ajudaram a acelerar a geração de dados. Há vários usos potenciais de *big data* para resolver problemas de

cidades, coletados diretamente da fonte ou por meio de análise de dados, inteligência de dados e mineração de dados (NUAIMI *et al.*, 2015).

Para Batty (2012), nossa obsessão atual com *big data* não se explicaria pelo tamanho dos conjuntos de dados, que tem aumentado continuamente no tempo, mas sim pela forma como os dados estão sendo coletados: muitos conjuntos de dados ainda são gerados por respostas humanas; entretanto uma proporção crescente está sendo produzida de forma automática e rotineira por sensores.

Inteligência em cidades está associada à geração automática, por sensores, de novos fluxos de dados georreferenciados, posteriormente armazenados em bases de dados a serem integradas para agregar valor. Esta integração é uma possibilidade duvidosa, que depende da existência de chaves comuns. Mesmo assim, estes dados que, até o momento, referem-se a movimentos e transporte, fluxos de energia e de outros insumos, além de dados espaciais sobre mercado de imóveis e vendas, são o cerne dos projetos de cidades inteligentes. Eles nos permitem entender como os sistemas urbanos funcionam no curto prazo (BATTY, 2012).

Os urbanistas da área de transportes sempre se preocuparam em entender como os sistemas funcionam no dia-a-dia, mas, na história do planejamento urbano, a maior parte da teoria e das aplicações têm funcionado na escala de anos: curto prazo de 5 anos e longo prazo de 20 a 50 anos. Cidades inteligentes podem significar uma mudança de ênfase do longo para o curto prazo. Na verdade, o que provavelmente acontecerá é que, se os dados forem coletados por um período longo, veremos no *big data* o tipo de mudança que antes só conseguíamos ver a cada dez anos nos censos completos (BATTY, 2012).

Essa visão otimista não é plenamente endossada por Bettencourt (2014) para quem questões econômicas e sociais em cidades são aquilo que os planejadores denominam "problemas perversos" (*wicked problems*), persistentes e não completamente explicitados. Revisitando os problemas perversos, sob a ótica da teoria da complexidade computacional, ele sustenta que o planejamento urbano seja intratável em termos computacionais. Isto significa que o conhecimento e a previsão de cadeias detalhadas de comportamentos têm a propriedade fundamental de serem praticamente impossíveis em cidades grandes, qualquer que seja a quantidade de dados disponíveis.

O conceito de problemas perversos foi criado por dois planejadores urbanos, Rittel e Webber (1973). A característica essencial dos problemas perversos é que eles não podem, na prática, ser resolvidos por um planejador central. Para melhor explicá-los, Bettencourt

(2014) utiliza o conceito de "Problema do Planejador", desmembrado em duas facetas distintas: (i) o problema do conhecimento; (ii) o problema do cálculo.

O problema do conhecimento refere-se à informação de que um planejador necessitaria para mapear e entender o estado atual do sistema, no nosso caso, a cidade. Se desconsiderarmos os problemas de privacidade, é concebível que as vidas e a infraestrutura física de uma grande cidade possam ser adequadamente monitoradas por sensores situados em milhões de locais, em taxas temporais finas, de modo a produzir taxas de informação elevadas, mas potencialmente gerenciáveis. Ou seja, o problema do conhecimento não é um obstáculo ao planejamento urbano (BETTENCOURT, 2014).

O problema do cálculo refere-se à complexidade computacional da tarefa de planejamento, em termos do número de passos necessários para avaliar todos os possíveis cenários e escolher o melhor curso de ação. A avaliação exaustiva de todos os possíveis cenários em uma cidade é impraticável, dado que envolve a consideração de enormes espaços de possibilidades (BETTENCOURT, 2014).

Bettencourt (2014) formaliza essa constatação por meio de um teorema que, considerando as possibilidades combinatórias, estabelece que a criação e a seleção do melhor e mais detalhado planejamento urbano requer uma computação envolvendo $O[P(N)]$ passos em uma cidade com população de tamanho N onde:

$$P(N) \simeq P_0 e^{N^{1+\delta} \ln(N)} \quad (19)$$

onde $\delta \approx \frac{1}{6}$ e P_0 é uma constante, independente de N . Para uma cidade com um milhão de habitantes isto corresponde a $P(N) = P_0 10^{6 \times 10^6}$, um número astronômico, muito maior que todos os átomos do Universo ($\approx 10^{82}$). Dessa forma, conclui-se que o Problema do Planejador seja praticamente impossível de resolver.

A aceitação dessa impossibilidade nos traz outra perspectiva sobre o uso de *big data* em cidades, diferente daquela adotada pela engenharia (usando teoria de sistemas e cibernética). Neste contexto, o papel principal do *big data* e das tecnologias de informação e comunicação em cidades seria facilitar os fluxos de informação e os mecanismos de aprendizado e coordenação entre indivíduos heterogêneos (BETTENCOURT, 2014).

No entanto, Bettencourt (2014) admite que, em certos casos, fontes maiores e mais amplas de dados possam realmente oferecer soluções radicalmente novas para problemas difíceis. Isto se deve ao fato dos eletrônicos modernos serem tão rápidos, se comparados à maioria dos fenômenos, físicos, biológicos e sociais, que vários problemas de política

urbana agora se enquadram nessa janela de oportunidade. Uma vantagem desta classe de problemas é que os modelos de resposta de sistema podem ser linearizados. Assim, a engenharia pode ignorar convenientemente as complexidades que surgem nestes sistemas em escalas temporais longas ou em escalas espaciais maiores, para as quais há necessidade de desenvolver modelos de comportamento humano. Muitos exemplos bem sucedidos de gestão e política urbanas baseadas em *big data* pertencem a essa categoria, a despeito de estarem sendo implementadas por pessoas, organizações, ou por algoritmos computacionais.

O quadro 1 mostra um sumário de importantes temas urbanos, suas escalas típicas, temporal e espacial, e a métrica dos resultados (*outcomes*). Nos sistemas de ônibus, os passageiros devem aguardar alguns minutos para serem transportados. Com base em medições do intervalo de tempo entre os ônibus em cada parada, e do número de passageiros que aguardam, o planejador pode usar uma solução baseada no controle de retroalimentação: comunicar-se com os ônibus para reforçar os padrões desejados de serviços, e colocar rapidamente menos ou mais unidades em serviço nos locais onde os parâmetros comecem a desviar-se da métrica ideal. Este tipo de estratégia pode ser operada intuitivamente por humanos ou por algoritmos que tenham acesso às medições e ações necessárias. Procedimentos similares podem ser adotados na gestão do fornecimento de água ou eletricidade. Gestão do tráfego, coleta de lixo, manutenção da infraestrutura urbana (edificações, estradas, dutos, etc) também podem ser gerenciados e integrados, de forma análoga. Em princípio, tampouco existe razão para que problemas sociais difíceis e persistentes, como o crime, não respondam ao patrulhamento ou a outras estratégias de aplicação da lei, como mostra a experiência bem sucedida de várias cidades, em especial, da cidade de Los Angeles, nos Estados Unidos. O crime é local e seus resultados bastante claros (BETTENCOURT, 2014).

Quadro 1 – Questões urbanas, suas escalas temporal e espacial e o caráter das métricas associadas

Problema	Escala Temporal	Escala Espacial	Métrica
Transporte(ônibus, metrô)	minutos	metros	simples
Fogo	minutos	metros	simples
Epidemias(HIV, Gripe)	anos, dias	tamanho da cidade	simples
Doenças Crônicas	décadas	tamanho da cidade	simples
Saneamento	anos	tamanho da cidade	simples
Crime	minutos	metros	simples
Infraestrutura	dias	metros	simples
Tráfego	minutos	metros -Km	simples
Coleta de lixo	dias	metros	simples
Educação	décadas	tamanho da cidade	complexa
Desenvolvimento econômico	décadas	tamanho da cidade	complexa
Emprego	anos	tamanho da cidade	complexa
Pobreza	décadas	vizinhança	complexa
Energia e Sustentabilidade	anos	tamanho da cidade	complexa
Habitação popular	anos-décadas	vizinhança	complexa

Fonte: Bettencourt (2014)

Em suma, para Bettencourt (2014), é possível identificar, usando *big data* e teoria de sistemas, soluções de transporte, fornecimento de água, energia, saneamento, coleta de lixo, controle de epidemias, e combate ao crime nas cidades. No entanto, ainda é difícil criar soluções desse tipo para problemas de educação, moradia, desenvolvimento econômico, pobreza ou sustentabilidade nas cidades. Para esses problemas, devido à sua escala temporal, a solução proposta é usar *big data* para subsidiar as decisões coletivas nas cidades. Vejamos exemplos dos dois tipos de problemas.

O combate aos crimes cibernéticos e aos crimes que deixam rastros na internet é um exemplo de uso do *big data* para revolucionar o tratamento de problemas sociais nas cidades, sem necessidade de desenvolver modelos de comportamento humano. Para Chauhan e Aluvalu (2016), os dados são semi-estruturados ou não estruturados, o que implica serem incompatíveis com métodos tradicionais de análise de dados. Nesse caso, a melhor solução passaria por métodos específicos de análise. Na primeira etapa, os dados primários coletados seriam distribuídos por localização geográfica e, com base nessa informação, seriam identificados *clusters*. A segunda etapa corresponderia à análise dos *clusters*. Ao final os *clusters* seriam processados por uma rede neural artificial, de modo a produzir uma padrão de previsão, a ser utilizado para combater o crime.

Recapitulando, dados digitais podem ser estruturados, semi-estruturados, ou não-estruturados. A maior parte dos dados digitais usados até o momento para prever e combater crimes são dados estruturados, ou seja, são organizados em formato tabular, com linhas e colunas. Por meio de técnicas tradicionais de aprendizado automático, como classificação e clusterização, os locais com maiores chances de ocorrência de crimes são identificados. No entanto, o uso da internet está crescendo rapidamente e é dessa forma que os criminosos se comunicam para realizar seu intento. O volume de dados gerados de forma automática na internet sobre esses crimes é enorme, mas seu formato é semi-estruturado e não estruturado. Sendo assim, Chauhan e Aluvalu (2016) acreditam que o uso de técnicas específicas de clusterização de *big data* associadas ao uso de redes neurais artificiais para previsão de padrões poderiam reduzir o tempo e aumentar a eficiência da investigação de crimes.

Passemos agora ao segundo tipo de problema, em que modelos de comportamento humano se fazem necessários para resolvê-los. Isso ocorre quando a escala temporal dos problemas urbanos é de anos ou décadas. Nesse caso, a abundância de dados apenas capacitaria pessoas e organizações a resolvê-los. Um exemplo clássico é o problema do desenvolvimento econômico.

Analisando a história econômica da Costa do Marfim, Andris e Bettencourt (2014) observaram que, naquele país, as pessoas com educação universitária ganhavam salários mais altos, mas estavam sujeitas a uma das mais altas taxas de desemprego e a longos períodos de busca por um emprego. Assim, a produção de uma população mais educada não era estatisticamente a resposta. Em vez disso, seria necessário entender como as inovações são aprendidas e usadas nas sociedades humanas e o papel das interações sociais nesses processos.

Nas nações em desenvolvimento, como a Costa do Marfim, o emprego informal domina a economia e está associado a baixos níveis de especialização e coordenação do trabalho, gerando baixo valor adicionado. Desse modo, a questão crucial torna-se: por que essas economias em desenvolvimento não adotam estratégias para promover a divisão e a interdependência do trabalho através de diferentes escalas de modo a promover melhores resultados econômicos e ciclos abertos de inovação tecnológica e organizacional? (ANDRIS; BETTENCOURT, 2014).

A resposta para a não realização desse intento pode estar na estrutura (incipiente) de seus sistemas urbanos. Para que uma nação se desenvolva é essencial que se crie

um sistema urbano constituído por grupos de cidades separadas, mas interdependentes, de modo que as vantagens das grandes cidades, em termos de inovação e organização, influenciem positivamente os setores primário e secundário da economia, predominantes nas cidades de pequeno porte (ANDRIS; BETTENCOURT, 2014).

Nas teorias das localidades centrais (Christaller (1966), Losch *et al.* (1954)), posteriormente elaboradas pela moderna geografia econômica (Fujita (2001)), a hierarquia funcional das atividades econômicas é a base para as dependências entre as cidades. Isso significa que a cidade de maior porte no sistema urbano contém todas as atividades econômicas observadas nas cidades menores, mas não o contrário. Assim, cidades de grande porte fornecem serviços (inovação, informação, organização) às pequenas cidades, em troca de comida e outros bens materiais. Apesar de algumas especializações urbanas, como as que ocorrem na manufatura de larga escala, essas tendências são características das nações mais desenvolvidas Mori, Nishikimi e Smith (2008).

Com base nessas premissas, Andris e Bettencourt (2014) buscaram entender em que medida o sistema urbano da Costa do Marfim estava espacialmente integrado e identificar os papéis das diferentes cidades à luz dessa hierarquia urbana nacional esperada. Para tanto, eles examinaram a estrutura das redes de chamadas telefônicas entre localidades (prefeituras) na Costa do Marfim.

O mapa do número total de chamadas, realizadas e recebidas, entre quaisquer duas localidades no sistema, mostrou, tal como esperado com base na teoria das localidades centrais, a importância da cidade de maior porte, Abidjan, com 3.8 milhões de habitantes. A conexão para Abidjan é expressa em termos da diversidade e da intensidade da troca de chamadas com muitas outras cidades da nação. A capital política Yamassoukro não pareceu desempenhar nenhum papel importante nessa rede, apesar da expectativa de que conectasse os dois maiores centros populacionais, Abidjan e Bouaké. Esses resultados mostraram que o sistema urbano costa-marfinense ainda é muito incipiente: muitas cidades mostram fortes ligações regionais, mas Abidjan e especialmente Yamoussoukro não conseguem manter uma rede de comunicação com a maior parte do país, especialmente com o Norte e Oeste, que são as partes mais pobres (ANDRIS; BETTENCOURT, 2014).

Cidades de grande porte são frequentemente descritas como máquinas, sociais, econômicas e políticas, para a maioria das nações desenvolvidas. Contudo, existem evidências de que a urbanização recente na África tenha falhado em promover o desenvolvimento econômico. A teoria de sistemas complexos e a economia urbana enfatizam

o papel das economias de aglomeração em todos esses processos; o resultado do processos sócioeconômicos aumenta *per capita* com o tamanho das cidades, tal como expresso na Lei de Bettencourt-West. Isso é interpretado em termos das possibilidades de interação criadas pela concentração espacial e temporal em cidades (BETTENCOURT, 2013).

Para avaliar em que medida a maior cidade da Costa do Marfim, Abidjan, cumpria o seu papel esperado, segundo essas teorias, Andris e Bettencourt (2014) analisaram a rede de chamadas telefônicas entre partes de Abidjan e áreas próximas. Os resultados obtidos mostraram que, embora a Costa do Marfim fosse relativamente desconectada como um sistema urbano, existia maior conectividade dentro das cidades, em especial, em Abidjan, o que teria propiciado a geração de economias de aglomeração. De fatos, os índices de pobreza em Abidjan são menores que em outras partes do país, especialmente se comparados aos índices das áreas rurais.

Os achados desse estudo corroboram a tese de que a chave para o desenvolvimento econômico seja o acesso mais rápido à participação cívica e econômica de populações pobres e excluídas. Esse seria o papel das Tecnologias de Informação e Comunicação (TIC) e do *big data* na promoção do desenvolvimento econômico (BETTENCOURT, 2014).

4 Metodologia da pesquisa

A Metodologia da pesquisa está organizada da seguinte forma.

Antes de mais nada, justificaremos a opção pelos municípios, em vez das RMs, como entes equivalentes às cidades, tal como conceituadas no Referencial Teórico.

Em seguida, apresentaremos o Plano de Dados Numéricos e Textuais que abrange duas fontes de dados: DOU e SIOP.

Logo após haverá uma breve introdução ao Processamento de Linguagem Natural, em especial, aos algoritmos a serem utilizados na extração de dados do DOU, o que corresponde ao primeiro Objetivo Específico.

Ao final descreveremos as ferramentas estatísticas utilizadas para identificação de padrões de escalamento das transferências de recursos, da União Federal para os municípios brasileiros, na modalidade saneamento básico, em relação à população destes municípios, tal como estabelecido no segundo Objetivo Específico.

4.1 *Municípios e regiões metropolitanas*

Os municípios brasileiros são uma circunscrição territorial dotada de personalidade jurídica e com autonomia política e administrativa. Uma peculiaridade do caso brasileiro é que os municípios são entidades estatais integrantes da Federação, dotadas de autonomia política, administrativa e financeira (CORRALO, 2006).

Atualmente há 5.570 municípios distribuídos pelo território nacional. Em geral, esses municípios são constituídos por várias cidades ou distritos, entremeados por áreas rurais conectadas, econômica e socialmente, à cidade principal. A cidade principal é denominada distrito-sede, enquanto as menores não possuem autonomia municipal, ou seja, não são emancipadas.

Dados divulgados pelo IBGE, referentes ao ano de 2018, mostram que juntos, as 26 capitais e o Distrito Federal, abrigam 49,7 milhões de habitantes, ou 23,8% da população do País. Dentre os 17 municípios com mais de um milhão de habitantes, 12 são capitais. Os três maiores centros urbanos são: (i) São Paulo - 12,2 milhões; (ii) Rio de Janeiro - 6,7 milhões; e (iii) Brasília - 3 milhões. Em contraposição, há três municípios com menos de

mil habitantes, os menos populosos do País: (i) Serra da Saudade/MG- 786; (ii) Borá/SP- 836; e (iii) Araguainha/MT - 956.

Alguns municípios brasileiros integram aglomerados urbanos classificados como RMs. A tabela 5 classifica os municípios metropolitanos por faixas populacionais.

Tabela 5 – Municípios metropolitanos e população, segundo faixas populacionais - 2016

Faixas	Número de Municípios	População	%
Até 100 mil	997	20.746.700	18,4
101 mil a 500 mil	148	32.690.385	29,0
501 mil a 1 milhão	21	14.257.220	12,6
1,001 a 2 milhões	10	13.554.439	12,0
2,001 a 3 milhões	5	13.132.866	11,6
Mais que 3 milhões	2	18.537.012	16,4
TOTAL	1.183	112.927.652	100,0

Fonte: Marguti, Costa e Favarão (2018)

Não encontramos uma definição única mas, com base na literatura e em textos legais, acreditamos ser possível afirmar que uma RM seja o conjunto de diferentes cidades próximas e interligadas, normalmente construído ao redor de uma metrópole, uma cidade central e mais desenvolvida. Geralmente, elas formam aglomerações urbanas, grandes áreas urbanizadas constituídas pela cidade núcleo e cidades adjacentes, onde são observadas conurbações, ou seja, o desaparecimento dos limites físicos entre as cidades. No entanto, RMs não precisam ser obrigatoriamente constituídas por uma única área contígua urbanizada, podendo ser também constituídas por uma região com duas ou mais áreas urbanizadas, intercaladas com áreas rurais. O fator preponderante para a definição de RM é a existência de um alto grau de integração, econômica, política e cultural, entre as cidades que formam o aglomerado urbano.

As primeiras delimitações de RMs ocorreram na década de 1960, com base em estudo do IBGE, a partir da aplicação de critérios aos dados do Censo de 1970, orientada por métodos estatísticos e noções teóricas sobre urbanização e metropolização. Com base nesse estudo, foram criadas as 9 primeiras RM: São Paulo, Belo Horizonte, Porto Alegre, Curitiba, Salvador, Recife, Fortaleza, Belém, e Rio de Janeiro (CUNHA, 2015).

Posteriormente, a CF de 1988 facultou aos estados brasileiros instituir regiões metropolitanas, aglomerações urbanas e microrregiões, constituídas por agrupamentos de municípios limítrofes, para integrar a organização, o planejamento e a execução de funções públicas de interesse comum (BRASIL, 1998; CUNHA, 2015).

Sendo assim, podemos classificar as RMs brasileiras em dois grupos: (i) criadas pelo governo federal, antes da CF de 1988; e (ii) criadas pelos governos estaduais, após a CF de 1988. Tal divisão explica o porquê de tratamentos bem diferentes para temas como estrutura, governança, gestão, financiamento e atuação, além do fato de que, após 1988, o novo quadro federativo brasileiro garantiu autonomia aos municípios, dificultando o papel integrador e planejador dos governos estaduais (MARGUTI; COSTA, 2016; MARGUTI; COSTA; FAVARÃO, 2018).

No primeiro grupo, verifica-se uma desigualdade de situações, resultando em maior ou menor sofisticação das estruturas de governança, em maior ou menor controle social, embora quase todas RMs do grupo possuam uma estrutura de financiamento com agências e fundos de desenvolvimento metropolitano (MARGUTI; COSTA, 2016; MARGUTI; COSTA; FAVARÃO, 2018).

No segundo grupo, observa-se, embora algumas tenham sido criadas nos anos 1990, que o surgimento de novas RMs se intensificou nos anos 2000 – quando ser RM virou critério de programas federais com grandes montantes de recursos para infraestrutura. Isso se tornaria ainda mais evidente a partir de 2009, com a criação do programa federal “Minha Casa Minha Vida”, resultando na ampliação do processo de institucionalização de novas RMs no País, sem que houvesse um critério claro para além do fato de que podiam ser criadas pelos governos estaduais (MARGUTI; COSTA, 2016; MARGUTI; COSTA; FAVARÃO, 2018).

Em janeiro de 2015, foi instituído o Estatuto da Metrópole, ou Lei nº 13.089, que manteve os governos estaduais como responsáveis pela instituição das RMs, definidas como “uma aglomeração urbana que configure uma metrópole” (CUNHA, 2015).

A despeito das expectativas de que o Estatuto da Metrópole reduzisse a fragmentação institucional da gestão metropolitana, não há evidências de que isso tenha ocorrido. As leis específicas das RMs continuam se mostrando muito variadas e limitadas no tocante à definição dos principais temas e aspectos associados à questão metropolitana. Poucas estabelecem a criação de um sistema de gestão adequado à complexidade da gestão compartilhada ou instituem conselhos específicos e, em sua maioria, são omissas quanto ao financiamento do desenvolvimento metropolitano (MARGUTI; COSTA, 2016; MARGUTI; COSTA; FAVARÃO, 2018).

A RM de São Paulo, a maior em termos populacionais, possui mais de 21 milhões de habitantes, seguida pela do Rio de Janeiro, com 12 milhões. A tabela 6 traz o número de RMs, a população estimada, e o número de municípios para as 83 RMs do País.

Tabela 6 – RMs, população, porcentagem e número de municípios por RM, segundo faixas populacionais - 2016

Faixas	Nº RMs	População	%	Nº Municípios
Até 100 mil	4	232.092	0,2	20
101 mil a 500 mil	37	9.199.967	8,1	460
501 mil a 1 milhão	18	12.288.464	10,9	215
1,001 a 2 milhões	9	13.312.790	11,8	140
2,001 a 3 milhões	5	11.994.890	10,6	106
3,001 a 4 milhões	5	18.480.128	16,7	91
4,001 a 5 milhões	2	8.568.052	7,6	57
Mais que 5 milhões	3	38.851.269	34,4	94
TOTAL	83	112.927.652	100,0	1.183

Fonte: Marguti, Costa e Favarão (2018)

É notável que nada menos que 34,4% da população residente em RMs concentre-se em apenas três: Belo Horizonte, Rio de Janeiro e São Paulo, as maiores concentrações urbanas brasileiras, abrangendo 94 municípios. No outro extremo, temos quatro RMs com 20 municípios, que possuem apenas 0,2% da população em áreas metropolitanas oficiais: Central, Sul de Roraima (ambas em Roraima), Araruna e Barra de Santa Rosa (ambas na Paraíba). A maioria das RMs encontra-se na faixa populacional de 101 mil até 500 mil habitantes, totalizando 460 municípios, que, entretanto, concentram apenas 8,1% da população total (MARGUTI; COSTA; FAVARÃO, 2018).

Passemos agora à avaliação sobre se o conceito de cidade do Referencial Teórico é mais compatível com o conceito de município ou de RM.

No Referencial Teórico, vimos que cidades foram reconhecidas como formas fractais por arquitetos, planejadores e outros, mesmo antes do conceito de fractal ser formalmente formulado, embora a adaptação do conceito matemático a aplicações do mundo real não seja exata nem desprovida de complicações. Como ambas as definições, de município e de RM, admitem a existência de espaços vazios entre as áreas contínuas, e a repetição de padrões espaciais em diversas escalas (cidades maiores e menores), julgamos que ambos se assemelhem a formas fractais, sendo a RM uma espécie de município em escala maior.

Passemos então à verificação de outras possíveis diferenças entre município e RM.

Conforme mencionado anteriormente, a autonomia política e administrativa das RMs varia caso a caso, sendo algumas institucionalmente mais consolidadas que outras.

Em contraposição, nos termos da CF, o município brasileiro é entidade estatal integrante da federação, como entidade político-administrativa, dotada de autonomia política, administrativa e financeira. Ou seja, os municípios estão institucionalmente consolidados.

Ademais o marco legal do saneamento básico, instituído em 2007, estabelece que a titularidade dos serviços de saneamento básico pertence aos municípios (e não aos estados). Porém, quanto às Regiões Metropolitanas, essa titularidade não está definida de forma consolidada, em razão de disputas judiciais entre estados e municípios.

Por essa razão, de ordem jurídico-institucional, e também porque as RMs existentes, ao contrário dos municípios, não cobrem todo o território brasileiro, essa pesquisa analisará somente as transferências de recursos para saneamento básico, da União Federal para os municípios, e não para as RMs.

4.2 *Plano de dados numéricos e textuais*

4.2.1 Estrutura e acesso ao DOU

O DOU é um periódico mantido pela União Federal para publicação de atos oficiais, divulgado em formato PDF (*Portable Document Format - PDF*), por meio do Portal da Imprensa Nacional. As edições posteriores a 30/11/2017 também estão disponíveis no formato HTML (*Hypertext Markup Language - HTML*).

Os textos são divulgados em formato digital, com frequência mínima diária. Futuramente pretende-se que sejam gerados de forma automática, sempre que transferências de recursos ou quaisquer outras operações financeiras forem realizadas pela União Federal.

Conforme disposto no artigo 2º da Portaria nº 268/2009, da Imprensa Nacional, o DOU é dividido em três Seções.

Na Seção 1 são publicados os "Atos Normativos de interesse geral", na forma de leis, decretos, resoluções, instruções normativas, portarias e outros. Na Seção 2 são publicados os "Atos de Pessoal relativos aos servidores públicos". Na Seção 3 são publicados "extratos de instrumentos contratuais" (acordos, ajustes, autorizações de compra, contratos, convênios, ordens de execução de serviço, termos aditivos e instrumentos congêneres), editais de citação, intimação, notificação e concursos públicos, comunicados, avisos de licitação entre outros atos da administração pública decorrentes de disposição legal. Portanto, na Seção 3

estão os documentos que retratam as transferências de recursos da União Federal para os municípios.

O Portal da Imprensa Nacional oferece duas ferramentas de acesso ao DOU em formato PDF. Ambas são ferramentas de acesso gratuito e, por esse motivo, iremos utilizá-las. Não exploramos a possibilidade de obter acesso ao DOU mediante assinatura paga.

A primeira é a Pesquisa Avançada, que permite buscar termos ou palavras-chave, em determinado período de tempo, nas três Seções do DOU, desde o ano 2000 até os dias atuais. A busca pode seguir, por exemplo, os seguintes critérios: "Seção 3", "extrato de convênio", e "01/08/2018 a 17/08/2018". Se buscarmos a expressão "abastecimento de água", segundo esses critérios, obteremos vários Extratos de Convênio, referentes a projetos de abastecimento de água no período desejado.

A figura 18 mostra um típico documento obtido por meio da Pesquisa Avançada. Trata-se de um Extrato de Convênio, publicado em 01/08/2018, tendo por objeto a Implantação de Sistemas de Abastecimento de Água, como conveniente o Município de Parnaíba-PI, e concedente a Companhia de Desenvolvimento do Vale do São Francisco, vinculada ao então Ministério da Integração Regional. É possível verificar a autenticidade do documento por meio de um *link* que remete para a página do DOU da qual ele foi extraído.

Figura 18 – Extrato de convênio, 01/08/2018, município de Parnaíba- PI

EXTRATO DE CONVÊNIO

Espécie: Convênio Nº 865462/2018, Nº Processo: 59570000146201871, Concedente: COMPANHIA DE DESENV. DO VALE DO SAO FRANCISCO, Conveniente: MUNICIPIO DE PARNAIBA CNPJ nº 06554430000131, Objeto: Implantação de Sistemas de Abastecimento de Água no município de Parnaíba - PI., Valor Total: R\$ 350.400,00, Valor de Contrapartida: R\$ 400,00, Valor a ser transferido ou descentralizado por exercício: 2018 - R\$ 350.000,00, Crédito Orçamentário: Num Empenho: 2018NE800003, Valor: R\$ 350.000,00, PTRES: 128414, Fonte Recurso: 0263532010, ND: 44404233, Vigência: 24/07/2018 a 24/07/2020, Data de Assinatura: 24/07/2018, Signatários: Concedente: INALDO PEREIRA GUERRA NETO CPF nº 882.102.004-53, Conveniente: FRANCISCO DE ASSIS DE MORAES SOUZA CPF nº 010.900.463-91.

Este conteúdo não substitui o publicado na versão certificada (pdf).

Fonte: Brasil (2018a)

A segunda ferramenta é a Leitura de Jornais, disponível para todas as edições regulares e extras, de 2017 a 2019. O acesso gratuito ao conteúdo completo da edição do dia somente está liberado somente em períodos restritos, de menor tráfego no sítio da Imprensa Nacional.

Utilizando a Leitura de Jornais, selecionamos, em primeiro lugar, ano, dia, e mês, por exemplo, 22/08/2018, e, em seguida, a Seção 3 regular. Se não houver conteúdos disponíveis na seção e dia selecionados, haverá um aviso de indisponibilidade. Caso contrário, o conteúdo virá classificado por tipo de documento e por órgão. A figura 19 mostra uma captura de tela com o resultado da pesquisa baseada nestes critérios, a qual mostra vários tipos de documentos referentes aos atos do Ministério das Cidades.

As duas ferramentas facilitam a busca, por órgão e por tipo de documento, mas dificultam a obtenção dos montantes totais transferidos pela União Federal para os municípios, na modalidade saneamento básico, referentes a um determinado período, como pretendíamos. Para tanto, seria necessário extrair as informações desejadas, documento por documento, órgão por órgão, no período desejado, tarefa demorada e complicada.

Figura 19 – Leitura de jornais, 22/08/2018, Seção 3, MCidades

Ministério das Cidades

Companhia Brasileira de Trens Urbanos

Administração Central

[AVISO DE LICITAÇÃO RDC ELETRÔNICO Nº 2/2018 - UASG 275068](#)

[EXTRATO DE CONTRATO](#)

[EXTRATO DE TERMO ADITIVO](#)

[EXTRATO DE TERMO ADITIVO](#)

Superintendência de Trens Urbanos de Belo Horizonte

[EXTRATO DE TERMO DE CONFISSÃO DE DÍVIDA](#)

Superintendência de Trens Urbanos de Maceió

[AVISO DE HOMOLOGACAO PREGAO ELETRONICO Nº 20/2017/COLIC/STU-MAC/CBTU](#)

[resultado DE HABILITAÇÃO TOMADA DE PREÇOS Nº 1/2018/COLIC/STU-MAC/CBTU](#)

Superintendência de Trens Urbanos de Recife

[AVISO DE INEXIGIBILIDADE de licitação Nº 4/GOLIC/2018](#)

[EXTRATO DE TERMO ADITIVO](#)

Empresa de Trens Urbanos de Porto Alegre S/A

[AVISO DE INEXIGIBILIDADE DE LICITAÇÃO](#)

[AVISO DE LICITAÇÃOOPREGÃO ELETRÔNICO SRP Nº 60/2018](#)

[AVISO DE LICITAÇÃOOPREGÃO ELETRÔNICO SRP Nº 81/2018](#)

[AVISO DE LICITAÇÃOOPREGÃO ELETRÔNICO SRP Nº 83/2018](#)

[AVISO DE PREÇOS REGISTRADOS SRP Nº 42/2018](#)

[EDITAL Nº 57, DE 12 DE ABRIL DE 2018CONVOCAÇÃO DE CANDIDATO](#)

Secretaria Nacional de Habitação

[EXTRATO DE ADESAO](#)

Fonte: Brasil (2018a)

4.2.2 Estrutura do SIOP

O SIOP é o sistema informatizado que suporta os processos de planejamento e orçamento da União Federal. O sistema começou a ser desenvolvido, em 2009, em substituição ao antigo Sistema Integrado de Dados Orçamentários (SIDOR).

Analisando a figura 20, que mostra o conjunto de sistemas que suportam os processos de planejamento, orçamento e execução da União Federal, pode-se inferir que o SIOP corresponde às antigas funções do SIDOR e do Sistema de Informações Gerenciais e de Planejamento do Plano Plurianual (SIGPLAN).

Figura 20 – Famílias de sistemas de apoio da Administração Pública Federal

Fonte: SIOPDoc... (2018)

O SIOP trata dos seguintes processos:

- Elaboração e revisão do Projeto de Lei do Plano Plurianual;
- Elaboração do Projeto de Lei de Diretrizes Orçamentárias;
- Elaboração do Projeto de Lei Orçamentária Anual;
- Alterações Orçamentárias/Créditos;
- Orçamento Impositivo;
- Receitas;
- Acompanhamento das Estatais;

- Acompanhamento Orçamentário; e
- Monitoramento do PPA.

Essas funcionalidades estão organizadas nos vários módulos do SIOP que também possui ferramentas de consulta e análise denominadas SIOP-Gerencial ou Consultas SIOP.

O Painel do Orçamento é uma ferramenta do SIOP para acesso livre e gratuito aos dados da LOA. Qualquer cidadão com acesso à internet pode obter, por meio desta ferramenta informações atualizadas sobre orçamento, sem necessidade de autenticação, nem mesmo um cadastro prévio. O endereço www.siop.planejamento.gov.br/siop possui um link para o Painel do Orçamento.

Os dados do SIOP não abrangem recursos de natureza extra-orçamentária, tais como o FGTS, cujo agente financeiro é a CEF, ou o FAT, gerido pelo BNDES, tampouco os recursos próprios de outros agentes financeiros da área de saneamento básico. O que torna imprescindível a consulta direta a outros documentos públicos, caso se queira abarcar todas as transferências da União Federal para os municípios, na modalidade de saneamento básico.

Não é necessário saber utilizar ferramentas de Linguagem de Consulta Estruturada (*Structured Query Language- SQL*) para extrair os dados, pois a interface é amigável o suficiente para ser utilizada por qualquer pessoa com noções básicas sobre orçamento público.

Na atual estrutura, as programações orçamentárias estão organizadas em programas de trabalho, que contêm informações qualitativas e quantitativas, físicas e financeiras (BRASIL, 2018c).

Do ponto de vista qualitativo, o programa de trabalho deve responder às perguntas clássicas sobre orçamento, sendo composto pelos seguintes blocos de informação (BRASIL, 2018c):

- Esfera Orçamentária;
- Classificação Institucional;
- Classificação Funcional;
- Estrutura Programática;
- informações Principais do Programa; e
- Informações Principais da Ação.

A Classificação por Esfera Orçamentária permite identificar a qual dos orçamentos definidos no artigo 165, parágrafo quinto da CF, uma determinada programação está associada (BRASIL, 2018c):

- 10 - Orçamento Fiscal;
- 20 - Orçamento da Seguridade Social; e
- 30 - Orçamento de Investimento.

A Classificação Institucional permite identificar o programa de trabalho do governo segundo os Poderes, Órgãos e Unidades Orçamentárias (UO). O objetivo do classificador é identificar a responsabilidade institucional pelo gasto. Expresso por um código numérico de cinco dígitos, esse classificador pode ser assim interpretado: o primeiro dígito identifica o Poder; o segundo o Órgão; e os três últimos designam a UO responsável pelas ações orçamentárias (BRASIL, 2018c).

Um órgão orçamentário ou uma UO não correspondem necessariamente a uma estrutura administrativa. Como exemplos podemos citar alguns fundos especiais e os órgãos "Transferências a Estados, Distrito Federal e Municípios", "Encargos Financeiros da União", "Operações Oficiais de Crédito", "Refinanciamento da Dívida Pública Mobiliária Federal" e "Reserva de Contingência" (BRASIL, 2018c).

A Classificação Funcional permite agregar os gastos do governo em grandes áreas de atuação governamental tais como saneamento básico. Também expresso por um código de cinco dígitos, esse classificador pode ser assim interpretado: os dois primeiros dígitos dizem respeito à função, que guarda relação com a missão institucional do órgão ou instituição; os três dígitos seguintes referem-se à subfunção, que pode ser entendida como uma partição da função. Por exemplo, dentro da função "Saneamento Básico" temos a subfunção "Saneamento Básico Urbano" (BRASIL, 2018c).

Toda ação na esfera federal está estruturada em programas orientados para a realização dos objetivos estratégicos definidos para o período do Plano Plurianual (PPA), ou seja, quatro anos. O PPA contempla dois tipos de programa: Programa Temático, definido como aquele que expressa e orienta a ação governamental para a entrega de bens e serviços à sociedade; e Programa de Gestão, Manutenção e Serviços ao Estado: aquele que expressa e orienta as ações destinadas ao apoio, à gestão e à manutenção da atuação governamental. Esses programas também constam da LOA. Todavia, na LOA há alguns programas que não constam no PPA, que são os programas compostos por Operações

Especiais, tais como o Serviço da Dívida Externa. No SIOP, o campo que identifica o programa contém quatro dígitos (BRASIL, 2018c).

A cada programa, estão associados um objetivo e um conjunto de ações orçamentárias que podem ser definidas como operações das quais resultam produtos (bens ou serviços) que contribuem para atender ao objetivo de um programa. No SIOP, cada ação orçamentária é identificada por um código alfanumérico de quatro dígitos, acrescido de quatro dígitos do localizador da ação (BRASIL, 2018c).

Dentre os vários atributos das ações orçamentárias, destaca-se a forma de implementação, a qual pode ser classificada como (BRASIL, 2018c):

- implementação direta;
- implementação descentralizada/delegada
- transferências obrigatórias e outras; e
- linha de crédito.

Na implementação direta, a ação orçamentária é executada diretamente pela unidade responsável, sem que ocorra transferência de recursos financeiros para outros entes da Federação (Estados, Distrito Federal e Municípios) ou ainda para entidades privadas (BRASIL, 2018c).

Como exemplo de implementação direta, vimos que, em casos excepcionais, as ações a cargo da FUNASA podem ser executadas diretamente pelo órgão.

Na implementação descentralizada ou delegada, temos atividades ou projetos, que estejam na área de competência da União, e que sejam executadas por outro ente da Federação, com recursos repassados pela União (BRASIL, 2018c).

As ações executadas com recursos não onerosos, que não mantenham vinculação com o Programa de Aceleração do Crescimento (PAC), são usualmente enquadradas como implementação descentralizada ou delegada. Os instrumentos jurídicos para a transferência dos recursos podem ser convênios, contratos de repasse, termos de cooperação ou congêneres.

Por sua vez, as transferências obrigatórias são definidas como operações especiais que transferem recursos, por determinação constitucional ou legal, aos Estados, Distrito Federal e Municípios.

Por determinação da Lei nº 11.578/2007, a execução das ações do PAC, pelos órgãos e entidades dos Estados, Distrito Federal e Municípios, com o apoio da União Federal, é

viabilizada por meio de transferência obrigatória. O instrumento jurídico firmado entre a União e os entes federados proponentes é o Termo de Compromisso.

As transferências classificadas como outras são aquelas destinadas a entidades públicas ou provadas sem fins lucrativos, organizações não governamentais e outras instituições, que não decorram de determinação constitucional ou legal (BRASIL, 2018c).

Finalmente, a linha de crédito é uma ação realizada mediante empréstimo de recursos aos beneficiários da operação. Enquadram-se nessa classificação os casos de empréstimos concedidos por estabelecimento oficial de crédito a estados e Distrito Federal, municípios e ao setor privado (BRASIL, 2018c).

4.2.3 Extração de dados do SIOP

Passamos agora a descrever os critérios adotados e as etapas de extração de dados que foram realizadas para atender ao segundo Objetivo Específico da pesquisa.

Inicialmente, foram selecionadas, para o período 2011 a 2018, as ações orçamentárias vinculadas aos seguintes programas do PPA, identificados por Brasil (2018b) como saneamento básico:

- Saneamento Básico(2068);
- Planejamento Urbano(2054);
- Segurança Alimentar e Nutricional(2069);
- Oferta de Água(2051);
- Gestão de Riscos e Resposta a Desastres(2040);
- Serviços Urbanos de Água e Esgoto(0122);
- Gestão da Política de Desenvolvimento Urbano(0310);
- Drenagem Urbana e Controle de Erosão Marítima e Fluvial(1138);
- Urbanização, Regularização Fundiária e Integração de Assentamentos Precários (1128);
- Proteção e Promoção dos Direitos dos Povos Indígenas(2065);
- Resíduos Sólidos(2067);
- Saneamento Rural(1287);
- Infraestrutura Hídrica(0515);
- Conservação e Gestão de Recursos Hídricos(2026);

- Apoio ao Desenvolvimento Urbano de Municípios de Pequeno Porte(6001);
- Fortalecimento da Gestão Urbana(1136);
- Resíduos Sólidos Urbanos(80087);
- Apoio ao Desenvolvimento Urbano de Municípios de Médio e Grande Porte(6002);
- Integração de Bacias Hidrográficas(1036); e
- Revitalização de Bacias Hidrográficas em Situação de Vulnerabilidade e Degradação Ambiental(1035)

Para cada ação orçamentária desses programas, foram utilizados os filtros "modalidade de aplicação" e "localizador do gasto".

A modalidade de aplicação discrimina se os recursos financeiros foram aplicados diretamente por órgãos ou entidades no âmbito da mesma esfera de governo ou por outra esfera de governo (estados, DF e municípios), visando impedir contagens em dobro. Por meio desse filtro, foi possível contabilizar somente as ações orçamentárias referentes às transferências da União Federal para os municípios, excluindo-se as transferências para estados, regiões metropolitanas e ações executadas diretamente pela União.

O localizador do gasto refere-se à localização geográfica do gasto, identificada por um código numérico de quatro posições. Por meio deste filtro foi possível excluir os localizadores "Nacional" e "Estados".

Combinando-se os dois filtros, obtivemos as transferências da União Federal para os municípios, na modalidade saneamento básico, executadas nos próprios municípios.

A tabela 7 contém uma amostra não aleatória dos dados extraídos do SIOP, utilizando-se a metodologia anteriormente descrita. Para cada ação orçamentária foram capturadas informações sobre sua execução orçamentária e financeira, as quais ocorrem concomitantemente pois são vinculadas. Havendo orçamento mas não o financeiro não poderá ocorrer a despesa. Por outro lado, pode haver recurso financeiro, mas não se poderá gastá-lo, se não houver disponibilidade orçamentária. "Dotação Inicial" e "Dotação Atual" correspondem aos valores previstos na LOA e nos atos normativos que promovam a abertura de créditos adicionais. "Empenhado" corresponde aos valores previstos em ato emanado de autoridade competente que crie para o Estado a obrigação de pagamento, pendente ou não de implemento de condição. Trata-se de uma reserva de valores orçamentários para determinada ação.

Os dados populacionais são estimativas da população residente nos municípios, na data de referência de 01/07/2015, a cargo do Tribunal de Contas da União (TCU). A escolha das estimativas do TCU dentre outras opções, tais como os dados do IBGE, justifica-se pelo fato de uma parte expressiva das transferências da União Federal para os municípios, as denominadas transferências constitucionais, seguirem o critério populacional, sendo o TCU o órgão responsável por fornecer os dados sobre população para o cálculo dos montantes a serem destinados aos municípios.

A extração do SIOP, cujos resultados parciais estão na tabela 7, foi realizada em julho de 2018, e abrange informações sobre transferências de recursos, na modalidade saneamento básico, para 2.879 municípios, um pouco mais da metade (51,7%) dos 5.570 municípios brasileiros.

Muitas transferências foram planejadas, e seus valores registrados na coluna "Dotação Inicial", mas os recursos não foram posteriormente empenhados, e nesse caso o montante registrado na coluna "Empenhado" é igual a zero.

Para o ajuste da lei de escala, foram considerados os valores médios, transferidos pela União Federal diretamente aos municípios, na modalidade saneamento básico, no período de 2011 a 2018. A soma dos valores foi feita sem correção monetária das parcelas, dado que o objetivo da pesquisa é verificar uma proporção entre valores, expressa na lei de escala, e não sua evolução temporal, caso em que a correção dos valores torna-se necessária.

Tabela 7 – Transferências da União Federal para saneamento básico por municípios selecionados - 2011 a 2018 (em R\$1,00)

Municípios	População	Dotação Inicial	Dotação Atual	Empenhado	Liquidado	Pago
Serra da Saudade - MG	818	750.000	750.000	500.000	6900	6.900
Uru -SP	1.229	300.000	300.000	0	0	0
Montauri - RS	1.556	500.000	500.000	0	0	0
Contagem - MG	648.766	35.450.000	35.450.000	0	0	0
Nova Iguaçu - RJ	807.492	57.859.217	57.859.217	27.018.603	0	0
Duque de Caxias - RJ	882.729	4.850.000	4.850.000	3.300.000	0	0
Chapada da Areia - TO	1.397	200.000	200.000	0	0	0
Curitiba - PR	1.879.355	39.100.000	39.100.000	150.000	0	0
Manaus- AM	2.052.557	592.368.874	585.077.441	114.117.404	0	0
Macieira - SC	1.815	150.000	150.000	0	0	0
São José dos Campos- SP	688.597	57.580.000	57.580.000	4.763.576	0	0
Osasco - SP	694.844	22.200.000	16.658.511	15.948.422	0	0
Flor do Sertão -SC	1.600	300.000	300.000	0	0	0
São Sebastião do Rio Preto - MG	1.601	300.000	300.000	0	0	0
Guabiju - RS	1.612	300.000	300.000	0	0	0
Ponte Branca - MT	1.618	500.000	500.000	0	0	0
Ribeirão Preto - SP	666.323	9.000.000	9.000.000	500.000	0	0
Presidente Castelo Branco - SC	1.650	300.000	300.000	0	0	0
Pouso Novo - RS	1.847	300.000	300.000	0	0	0
Aracaju - SE	632.744	317.612.508	317.612.508	79.045.573	0	0

Fonte: Brasil (2018h)

4.3 *Processamento de linguagem natural*

Esse item pretende ser uma breve introdução às ferramentas de Processamento de Linguagem Natural (*Natural Language Processing- NLP*), utilizadas nos *softwares* de extração de dados do DOU, previstos no primeiro Objetivo Específico.

A mineração de dados do DOU abrange não apenas a identificação dos nomes dos municípios beneficiários, mas também dos valores transferidos pela União Federal aos municípios, a fundo perdido ou por meio de operações de crédito aos operadores do sistema. O próximo passo é utilizar estes dados para identificar padrões de escalamento das transferências de recursos, em função do tamanho dos municípios, tal como consta no Segundo Objetivo Específico.

Dados são a matéria-prima dos processos de mineração. Eles podem se manifestar, basicamente, de duas formas: estruturada e não estruturada. Como exemplo de dados estruturados temos as bases de dados em sistemas informatizados convencionais. Tipicamente, essas bases são organizadas de forma que se tenham dados armazenados em estruturas tabulares, em que as linhas armazenam uma ocorrência de um evento caracterizado por um conjunto de colunas que representam características que descrevem um exemplar (instância) daquele evento. Na maioria dos casos, os dados estruturados são resultantes de processos de geração de dados inerentes a sistemas transacionais ou resultantes de observações e processos de medição. Esses dados geralmente são armazenados em um conjunto de tabelas relacionadas entre si (SILVA; PERES; BOSCARIOLI, 2017).

O DOU é um conjunto de textos em português brasileiro, ou seja, são dados não estruturados. Para fins de mineração, dados não estruturados precisam passar por uma etapa de pré-processamento, de forma que uma representação adequada lhes seja produzida (SILVA; PERES; BOSCARIOLI, 2017).

Na figura 21 vemos a arquitetura de um sistema simples de extração de informações, com suas etapas e respectivos formatos de dados de saída.

Figura 21 – Arquitetura de extração de informações

Fonte: Bird, Klein e Loper (2009)

4.3.1 Pré-processamento de dados textuais

As duas primeiras etapas mostradas na figura 21 referem-se ao pré-processamento de dados. Antes de executá-las, é necessário alterar o formato do texto, de Formato Portátil de Documento (Portable Document Format - PDF) para TXT, um arquivo de texto sem formatação.

Na primeira etapa, qual seja, a divisão do texto em sentenças, podemos usar, desde um mecanismo simples como considerar que a sentença deva ser quebrada após um ponto (.), até algo tão complexo quanto um classificador para identificar os limites da sentença. A biblioteca *Natural Language Toolkit - NLTK* do Python possui uma função padrão de divisão do texto em sentenças, pré-treinada em diversas línguas europeias, inclusive em português (BIRD; KLEIN; LOPER, 2009; HARDENIYA, 2015):


```
import nltk
from nltk import sent_tokenize
texto = "Olá, pessoal! Meu nome é Ludmila."
sentenças = sent_tokenize(texto, language = "portuguese")
print(sentenças)
['Olá, pessoal!', 'Meu nome é Ludmila']
```

A biblioteca NLTK também possui uma função padrão para divisão do texto em *tokens*, que geralmente são palavras ou pontuação (HARDENIYA, 2015):

```
from nltk.tokenize import word_tokenize
tokens= word_tokenize(texto, language= "portuguese")
print(tokens)
['olá', ',', 'pessoal', '!', 'Meu', 'nome', 'é', 'Ludmila', '.']
```

Termos irrelevantes são palavras muito frequentes que pouco ou nada contribuem para o significado de uma sentença. A remoção de termos irrelevantes é um procedimento usual em NLP mas, nesse caso, se removermos os termos irrelevantes, perderemos a visão do contexto, e os algoritmos empregados na terceira e quarta etapas da figura 21 necessitam do contexto para a extração de informações.

4.3.2 Etiquetagem morfossintática

Etiquetagem morfossintática (*Part of Speech Tagging - POS Tagging*) é o processo de conversão de uma sentença, na forma de uma lista de palavras, em uma lista de tuplas, onde cada tupla assume a forma (palavra, etiqueta). As etiquetas são classes gramaticais. As mais comuns são substantivo, verbo, adjetivo, artigo, e pronome. Ou seja, etiquetagem morfossintática é a classificação dos tokens de uma sentença em classes gramaticais. As sentenças são extraídas de uma base de dados textuais, denominada *corpus*, onde cada exemplar é denominado documento.

Muitos *corpora* contêm anotações linguísticas, tais como classes gramaticais, entidades nomeadas, estruturas sintáticas, dentre outras. Em português, os mais conhecidos são Tycho Brahe, Mac-Morpho e Bosque. Os dois primeiros são os maiores, com aproximadamente um milhão de *tokens*. Bosque tem cerca de 200.000, sendo constituído por textos em português brasileiro e europeu (PERKINS, 2014; FONSECA; ROSA; ALUÍSIO, 2015).

Para treinar os algoritmos de etiquetagem morfossintática, a serem analisados neste item, utilizamos o Mac-Morpho, o maior *corpus* etiquetado manualmente em português brasileiro, composto por 109 arquivos com textos extraídos do jornal "Folha de São Paulo". O quadro 2 mostra as principais etiquetas do Mac-Morpho.

Usando uma lista de expressões regulares é possível associar etiquetas aos *tokens*. Podemos supor, por exemplo, que qualquer palavra com o sufixo "ar" seja o infinitivo de um verbo e qualquer palavra com o sufixo "ação" seja um substantivo. Também é fácil definir quais palavras são pronomes ou artigos. Se aplicarmos essas regras às sentenças do corpus Mac-Morpho obteremos as classes gramaticais das palavras que compõem as sentenças (BIRD; KLEIN; LOPER, 2009):

definição de regras

```
patterns = [
    ( r'.*ar$', 'V'),
    ( r'.*ação$', 'N'),
    ( r'^(a|o|as|os)$', 'ART' ),
    ( r'^(em|por|de|da)$', 'PREP'),
]
```

aplicação das regras a uma sentença do corpus Mac-Morpho

```
regex_tagger = nltk.RegexpTagger(patterns)
from nltk.corpus import mac_morpho
mac_morpho_sents = mac_morpho.sents()
regex_tagger.tag(mac_morpho_sents[5])
[('Em', None), ('o', 'ART'), ('dia', None), ('15', None),
 ('', None), ('Dia', None), ('da', 'PREP'), ('Conservação', 'N'),
 ('do', None), ('Solo', None), ('', None), ('o', 'ART'),
 ('único', None), ('fato', None), ('a', 'ART'), ('festejar', 'V'),
 ('pode', None), ('ser', None), ('a', 'ART'), ('Convenção', None),
 ('Internacional', None), ('sobre', None), ('Desertificação', 'N')]
```

Quadro 2 – Conjunto de etiquetas do Mac-Morpho

Classe Gramatical	Etiqueta
Adjetivo	ADJ
Advérbio Conectivo Subordinativo	ADV-KS
Advérbio Relativo Subordinativo	ADV-KS-REL
Artigo (def ou indef)	ART
Conjunção Coordenativa	KC
Conjunção Subordinativa	KS
Interjeição	IN
Nome	N
Nome Próprio	NPROP
Numeral	NUM
Particípio	PCP
Palavra Denotativa	PDEN
Preposição	PREP
Pronome Adjetivo	PROADJ
Pronome Conectivo Subordinativo	PRO-KS
Pronome Pessoal	PROPESS
Pronome Relativo Conectivo Subordinativo	PRO-KS-REL
Pronome Substantivo	PROSUB
Verbo	V
Verbo Auxiliar	VAUX
Símbolo de Moeda Corrente	CUR

Fonte: USP (2003)

A etiquetagem morfossintática por meio de algoritmos é muito mais eficiente que o processo manual de criação de regras. É possível treinar um classificador, que nada mais é que um algoritmo de aprendizado de máquina, para descobrir quais sufixos de palavras são mais informativos. E então o classificador tomará a decisão sobre qual etiqueta atribuir a uma nova palavra ou conjunto de palavras, com base na informação sobre qual, dentre os sufixos mais comuns, a palavra ou conjunto de palavras possui (se possuir) (BIRD; KLEIN; LOPER, 2009).

Mas não precisamos nos restringir aos sufixos. Podemos definir uma função de extração de características que incorpore outras características da palavra, tais como extensão, número de sílabas ou ainda prefixos. Tampouco precisamos nos restringir às características das palavras. Podemos alterar a função de extração para acomodar características que dependam do contexto em que a palavra está inserida (BIRD; KLEIN; LOPER, 2009).

Se associarmos essa função a um classificador, e treinarmos o conjunto em um *corpus* de nosso interesse, obteremos um etiquetador morfossintático adaptado. Classificadores

Bayesianos Ingênuos e Classificadores de Entropia Máxima podem ser utilizados para tal fim.

O Classificador Bayesiano Ingênuo calcula, em primeiro lugar, a probabilidade *a priori* de cada etiqueta, verificando a sua frequência no conjunto de dados de treinamento. A contribuição de cada característica é então combinada à probabilidade *a priori*, de modo a obter uma estimação da verossimilhança de cada etiqueta (*label likelihood*). A etiqueta cuja verossimilhança estimada for a mais alta será então atribuída ao dado de treinamento. A probabilidade resultante pode ser entendida como uma estimação da probabilidade de que um dado, selecionado aleatoriamente no conjunto de treinamento, tenha ambos, a etiqueta atribuída e o conjunto de características, assumindo que as probabilidades das características sejam independentes. A premissa de independência facilita a combinação das diferentes características, pois não temos que nos preocupar com uma possível interação entre elas (BIRD; KLEIN; LOPER, 2009).

O Classificador de Entropia Máxima é similar ao Classificador Bayesiano Ingênuo. No entanto, em vez de usar probabilidades para determinar os parâmetros, utiliza técnicas de busca para encontrar o conjunto de parâmetros que maximize o seu desempenho. Especificamente, ele busca o conjunto de parâmetros que maximize a estimativa de verossimilhança total (*total likelihood*) definida como (BIRD; KLEIN; LOPER, 2009):

$$P(features) = \sum_{x \in corpus} P(label(x|features(x))) \quad (20)$$

A probabilidade de que um dado de treinamento (*input*) que tenha determinadas características receba uma dada etiqueta (*label*) é definida como:

$$P(label|features) = P(label, features) / \sum_{label} P(label, features) \quad (21)$$

Uma importante distinção entre os dois classificadores refere-se às questões que eles conseguem responder. Classificadores Bayesianos Ingênuos são classificadores generativos que geram modelos que predizem a probabilidade conjunta do par $P(input, label)$. Por sua vez, Classificadores de Entropia Máxima são exemplos de classificadores condicionais, que geram modelos que predizem $P(label|input)$, ou seja, a probabilidade de uma determinada etiqueta para um dado específico de treinamento. Em geral, modelos generativos são mais poderosos que modelos condicionais, visto que podemos calcular a probabilidade condicional $P(label|input)$, a partir da probabilidade conjunta $P(input, label)$, mas não o inverso (BIRD; KLEIN; LOPER, 2009).

Tomando como base esses dois classificadores, é possível desenvolver etiquetadores que examinem o contexto em que uma palavra aparece, bem como as suas características, a fim de determinar qual etiqueta lhe deve ser atribuída. Na biblioteca NLTK, há uma classe denominada `ClassifierBasedPOSTagger()` que utiliza classificadores para a etiquetagem morfo sintática. As características são extraídas das palavras, e então passadas por um classificador interno. O classificador processa as características e retorna uma etiqueta de classe gramatical (PERKINS, 2014).

A classe `ClassifierBasedPOSTagger()` é uma subclasse da `ClassifierBasedTagger()` que implementa um detector de características que localiza sufixos de múltiplos tamanhos, identifica algumas expressões, e utiliza etiquetadores Unigrama, Bigrama e Trigrama para produzir um conjunto de características para cada palavra. Os conjuntos de características assim produzidos são usados para treinar o classificador interno, que por sua vez é usado para etiquetar as palavras segundo a sua classe gramatical (PERKINS, 2014).

A implementação da classe para o Classificador Bayesiano Ingênuo, utilizando o *corpus* Mac-Morpho, resultou em acurácia nos dados de teste de aproximadamente 83,6%, superior à implementação da classe para o Classificador de Entropia Máxima cuja acurácia foi de aproximadamente 75% nos dados de teste. O código Python para treinamento dos dois etiquetadores está no Anexo A.

Antes de passarmos aos modelos sequenciais, vamos analisar os etiquetadores Unigrama, Bigrama e Trigrama, mencionados anteriormente. Usando o etiquetador Unigrama, associamos cada palavra à etiqueta mais provável para aquela palavra no *corpus*. Embora cada palavra possa assumir múltiplas etiquetas conforme o contexto, considera-se apenas a etiqueta mais frequente. Por exemplo, o etiquetador Unigrama irá atribuir a etiqueta "NPROP" a qualquer ocorrência da palavra "Salvador", desde que "Salvador" seja mais usada no *corpus* como nome próprio (NPROP) do que como adjetivo (ADJ). O treinamento do etiquetador Unigrama e posterior aplicação a outro texto estão mostrados no anexo A.

Um etiquetador N-grama é uma generalização do etiquetador Unigrama, cujo contexto é a palavra corrente em conjunto com os $N - 1$ *tokens* precedentes, portanto o N-grama seleciona a etiqueta mais provável no contexto dado. Ao combinarmos vários etiquetadores é possível aumentar a acurácia e também a eficiência do modelo. Assim, se o Bigrama atribuir a mesma etiqueta que o Unigrama em determinado contexto, o Bigrama descarta a instância de treinamento, mantendo o modelo Bigrama o menor possível. O mesmo se aplica aos etiquetadores Trigrama e Bigrama. No Anexo A mostramos o

treinamento da combinação dos três etiquetadores: Unigrama, Bigrama e Trigrama. Também incluímos o etiquetador *default* que atribui a etiqueta mais frequente no *corpus* a todas as palavras. No Mac-Morpho, a etiqueta mais frequente é N(substantivo). O modelo, cuja acurácia foi de cerca de 90% nos dados de teste, pode ser gravado, e depois recuperado para utilização em outro conjunto de dados, conforme consta no Anexo A.

Modelos Ocultos de Markov (*Hidden Markov Models - HMM*) também são utilizados como etiquetadores, para atribuir a sequência correta de etiquetas a uma sequência de dados ou para avaliar a probabilidade de um certa etiqueta ou de uma certa sequência de dados. O modelo HMM é uma extensão da cadeia de Markov, onde cada estado corresponde de forma determinística a um dado evento. No modelo HMM a observação é uma função de probabilidade do estado. O que ambos têm em comum é a premissa de que a probabilidade de transição de um estado para outro depende somente do estado atual - a série de estados que conduziram ao estado atual não é considerada. Ambos são também invariantes no tempo (JURAFSKY; MARTIN, 2014).

Uma cadeia de Markov é um processo estocástico que descreve uma sequência de eventos possíveis, na qual a probabilidade de um evento futuro depende apenas do estado atual. Podemos descrever formalmente uma cadeia de Markov como sendo uma espécie de modelo gráfico probabilístico: uma forma de representar premissas probabilísticas em um grafo. Uma cadeia de Markov é especificada pelos seguintes componentes (JURAFSKY; MARTIN, 2014):

1. $Q = q_1, q_2 \dots q_n$ - uma sequência de N estados;
2. $A = a_{11}a_{12} \dots a_{n1} \dots a_{nn}$ - uma matriz de probabilidades de transição A, onde cada a_{ij} é a probabilidade de transição do estado i para o estado j onde $\sum_{j=1}^n a_{ij} = 1 \forall i$;
3. q_0, q_f - um estado especial inicial e um estado final que não estão associados a observações.

Em uma cadeia de Markov de primeira ordem, a probabilidade de um determinado estado depende somente do estado anterior:

$$P(q_i | q_1 \dots q_{i-1}) = P(q_i | q_{i-1}) \quad (22)$$

A figura 22 mostra, à esquerda, uma cadeia de Markov para atribuir uma probabilidade a uma sequência(a) de eventos de tempo para a qual o vocabulário é constituído por HOT (muito quente), COLD (frio), and WARM (quente). À direita, vemos outra

cadeia de Markov que representa um modelo de uma sequência de duas palavras (bigrama). Usando esses modelos, podemos atribuir uma probabilidade a qualquer sequência de nosso vocabulário (JURAFSKY; MARTIN, 2014).

Figura 22 – Cadeias de Markov para tempo e palavras

Figure 9.1 A Markov chain for weather (a) and one for words (b). A Markov chain is specified by the structure, the transition between states, and the start and end states.

Fonte: Jurafsky e Martin (2014)

Cadeias de Markov são úteis para computar a probabilidade de uma sequência de eventos. Em muitos casos, no entanto, os eventos que nos interessam podem não ser diretamente observáveis. Por exemplo, não observamos as etiquetas ao analisarmos um texto; tudo o que vemos são palavras (e sequências de palavras) e temos que inferir as etiquetas corretas. Como elas não são observáveis, dizemos que estão ocultas. Um Modelo Oculto de Markov permite que lidemos com os eventos observáveis (palavras), e com os eventos ocultos (etiquetas morfossintáticas), que julgamos serem os fatores causais no nosso modelo probabilístico (JURAFSKY; MARTIN, 2014).

Podemos descrever formalmente o Modelo Oculto de Markov como um grafo direto ponderado probabilístico (os pesos das arestas representam a probabilidade de transição entre os estados inicial e final) especificado pelos seguintes componentes:

1. $Q = q_1, q_2 \dots q_n$ - uma sequência de N estados;
2. $A = a_{11}a_{12} \dots a_{n1} \dots a_{nn}$ - uma matriz de probabilidades de transição A , onde cada a_{ij} é a probabilidade de transição do estado i para o estado j onde $\sum_{j=1}^n a_{ij} = 1 \forall i$;
3. $O = o_1o_2 \dots o_T$, cada uma definida a partir de um vocabulário $V = v_1v_2 \dots, v_v$;
4. $B = b_i(o_T)$ - uma sequência de probabilidades de observação, também denominadas probabilidades de emissão, cada uma expressando a probabilidade de uma observação o_t ter sido gerada por um estado i ;

5. q_0, q_f - um estado especial inicial e um estado final que não estão associados a observações, junto com probabilidades de transição $a_{01}a_{02}...a_{0n}$ para fora do estado inicial e $a_{1F}a_{2F}...a_{nF}$ para dentro do estado final.

Um Modelo Oculto de Markov de primeira ordem adota duas premissas simplificadoras: (i) a probabilidade de um estado particular depende somente do estado anterior; e (ii) a probabilidade de observação de uma saída (*output observation*) depende somente do estado que produziu a observação q_i e não de quaisquer outros estados ou de quaisquer outras observações.

A figura 23 mostra um Modelo Oculto de Markov utilizado para relacionar o número de sorvetes consumidos por uma pessoa em determinado dia (as observações são retiradas do alfabeto $O = 1,2,3$) e as condições do tempo (HOT (quente) e COLD (frio), que são as variáveis ocultas).

Figura 23 – Modelo oculto de Markov para consumo de sorvetes e tempo

Fonte: Jurafsky e Martin (2014)

Quando usado como etiquetador, o modelo HMM retorna a sequência de etiquetas mais provável para uma dada sequência de palavras. No que difere de outras técnicas de etiquetagem, que frequentemente só etiquetam palavras isoladas, buscando otimizar cada etiquetagem individual, mas sem considerar a combinação ótima de sequências para unidades maiores, como as sentenças. Para tanto, HMM utiliza o algoritmo de Viterbi que computa de forma eficiente o caminho ótimo através do grafo, dada uma sequência de formas de palavras (JURAFSKY; MARTIN, 2014).

Tal como o modelo HMM, o etiquetador HMM baseia-se em duas premissas simplificadoras: (i) a probabilidade de aparecimento de uma palavra depende somente de sua

própria etiqueta e é independente das palavras e etiquetas vizinhas; e (ii) a probabilidade de uma etiqueta depende somente da etiqueta anterior, e não da sequência inteira de etiquetas (JURAFSKY; MARTIN, 2014).

Dessa forma, estabelecemos que a probabilidade de uma determinada sequência de etiquetas $(t_1...t_N)$ para uma dada sequência de palavras $(w_1...w_N)$ é proporcional ao produto $P(t_i|t_{i-1})P(w_i|t_i)$ (JURAFSKY; MARTIN, 2014):

$$P(tags|words) \approx P(t_i|t_{i-1})P(w_i|t_i) \quad (23)$$

Para encontrar a melhor sequência de etiquetas para uma dada sequência de palavras devemos maximizar $P(tags|words)$. Na prática, podemos fazer com que a probabilidade $P(tags|words)$ dependa não apenas da etiqueta anterior, mas sim de duas ou mais etiquetas anteriores. Para duas etiquetas fazemos (JURAFSKY; MARTIN, 2014):

$$P(tags|words) \approx P(t_{i-1}|t_{i-2})P(w_i|t_i) \quad (24)$$

Além do contexto ampliado (duas ou mais etiquetas), os etiquetadores HMM estado-da-arte incorporam outras características. Uma delas é permitir que o etiquetador conheça a localização do fim da sentença, introduzindo um marcador de fim de sentença para t_{n-1} onde t é a abreviação de *tag* ou etiqueta, em português.

Para explicar como funciona o etiquetador HMM, vamos calcular $P(w_i|t_i)$, a partir dos dados do *corpus* Mac-Morpho, e usando a Estimação de Máxima Verossimilhança (*Maximum Likelihood Estimation* - *MLE*). Inicialmente, vamos colocar etiquetas artificiais "START" e "END" respectivamente no início e final de cada sentença e computar as etiquetas com extensão até 5:

```
from nltk.corpus import mac_morpho
mac_morpho_tags_words = [ ]
for sent in mac_morpho.tagged_sents():
    mac_morpho_tags_words.append( ("START", "START") )
    mac_morpho_tags_words.extend([(tag[:5], word) for (word, tag) in sent])
    mac_morpho_tags_words.append( ("END", "END") )
```

Passemos então ao cálculo das Distribuições de Frequência Condicional e de Probabilidade Condicional e em seguida à estimativa da $P(t_i|t_{i-1})$.

```

cfid_tagwords = nltk.ConditionalFreqDist(mac_morpho_tags_words)
cpd_tagwords = nltk.ConditionalProbDist(cfid_tagwords, nltk.MLEProbDist)
mac_morpho_tags = [tag for (tag, word) in mac_morpho_tags_words ]
cfid_tags= nltk.ConditionalFreqDist(nltk.bigrams(mac_morpho_tags))
cpd_tags = nltk.ConditionalProbDist(cfid_tags, nltk.MLEProbDist)

```

Finalmente, vamos calcular a probabilidade da sequência de etiquetas 'START ART NPROP V PREP N END' para a sequência de palavras "O Município recebeu recursos para saneamento":

```

prob_tagsequence = cpd_tags["START"].prob("ART")*\
cpd_tagwords["ART"].prob("O")*\
cpd_tags["ART"].prob("NPROP") *cpd_tagwords["NPROP"].prob("Município")*\
cpd_tags["NPROP"].prob("V") *cpd_tagwords["V"].prob("recebeu") *\
cpd_tags["V"].prob("N") *cpd_tagwords["N"].prob("recursos") *\
cpd_tags["N"].prob("PREP")* cpd_tagwords["PREP"].prob("para") *\
cpd_tags["PREP"].prob ("N")*cpd_tagwords["N"].prob("saneamento")*\
cpd_tags["N"].prob("END")
print( prob_tagsequence)
6.341673459830365e-23

```

O método descrito permite determinar a probabilidade de uma única sequência de etiquetas. No entanto, para identificarmos a melhor sequência de etiquetas precisamos calcular a probabilidade de todas as sequências de etiquetas. O que o Algoritmo de Viterbi nos propicia é uma forma de computar todas as probabilidades, a mais rápida possível. Não vamos analisá-lo em detalhes, apenas ressaltar que se trata de um algoritmo de aprendizado supervisionado. A biblioteca NLTK fornece o pacote HMM que contém os algoritmos de Viterbi e de Baum Welch (aprendizado não-supervisionado).

Como o desempenho do HMM foi bastante inferior ao esperado, decidimos testar o etiquetador TnT (*Statistical Part-of-Speech - TnT*), baseado em modelos de Markov de segunda ordem que consideram triplas de palavras consecutivas, ou seja, trigramas. Em vez de construir uma cadeia *backoff* de unigramas, bigramas e trigramas, o etiquetador TnT considera todos os três modelos para escolher a melhor etiqueta. Ele tenta prever as etiquetas para a sentença inteira de uma só vez, escolhendo o melhor modelo para a

sentença inteira, com base nas probabilidades para cada etiqueta possível (BRANTS, 2000; HARDENIYA, 2015).

Para o treinamento do etiquetador TnT, também mostrado no Anexo A, conforme sugerido por Hardeniya (2015), alteramos a opção *default* $N = 1000$ para $N = 100$, onde N é o número de possíveis soluções que o etiquetador considera enquanto tenta prever as etiquetas para cada sentença. Reduzimos a acurácia em menos de 1%, mas em compensação reduzimos o tempo de treinamento. Também estabelecemos que as palavras desconhecidas deveriam receber a etiqueta mais frequente do Mac-Morho que é substantivo (N).

HMMs são modelos generativos que modelam a distribuição de probabilidade conjunta $P(X, Y)$ das observações e das sequências de etiquetas; os parâmetros são treinados para maximizar a probabilidade conjunta dos dados de treinamento. Para definir a probabilidade conjunta de observações e sequências de etiquetas, um modelo generativo precisa enumerar todas as possíveis sequências de observações, o que requer uma representação na qual as observações são entidades atômicas, tais como palavras. Em particular, não é prático representar múltiplas características que interagem ou dependências de longo alcance entre as observações, dado que o problema de inferência para esses modelos é intratável. Essa é uma das principais motivações para optarmos por modelos condicionais, em vez de modelos generativos (LAFFERTY; MCCALLUM; PEREIRA, 2001).

Modelos condicionais especificam as probabilidades de possíveis sequências de etiquetas, dada uma sequência de observações. A probabilidade condicional $P(X|Y)$ de uma sequência de etiquetas pode depender de características arbitrárias, e não independentes, da sequência de observações. As características escolhidas podem representar atributos de diferentes níveis de granularidade das observações (por exemplo, palavras e caracteres no texto). A probabilidade de uma transição entre etiquetas pode depender não somente da observação atual, mas também das observações passadas e futuras, quando disponíveis. Em contraposição, modelos generativos baseiam-se em premissas muito restritas de independência entre as observações (LAFFERTY; MCCALLUM; PEREIRA, 2001).

Modelos de Markov de Entropia Máxima (*Maximum Entropy Markov Models - MEMMs*) são modelos sequenciais condicionais que possuem todas as vantagens dos demais modelos condicionais, embora tenham uma vulnerabilidade que pode ser denominada "problema do viés da etiqueta": as transições que partem de um dado estado competem somente umas com as outras, em vez de competir com todas as outras transições existentes

no modelo. Por isso, os MEMMs têm sido preteridos em favor dos Campos Aleatórios Condicionais (*Conditional Random Fields- CRFs*) que possuem todas as vantagens dos modelos sequenciais condicionais, sem o problema do viés da etiqueta (LAFFERTY; MCCALLUM; PEREIRA, 2001).

Para treinar o etiquetador CRF, usamos a classe `CRFTagger()` que já leva em consideração algumas características da palavra ou *token* a ser etiquetado, tais como se começa com letra maiúscula ou não, sufixos, se é um número ou pontuação. A acurácia do `CRFTagger` foi superior a 92%, tal como pode ser verificado no Anexo A.

Até agora analisamos dois tipos de etiquetadores: baseados em regras; e baseados em modelos estatísticos. Uma terceira e relativamente nova abordagem é a etiquetagem por redes neurais artificiais. Redes neurais são compostas por um grande número de unidades simples de processamento, denominadas neurônios, interconectadas por arestas ponderadas. Associada à cada unidade existe uma unidade de ativação. Por meio das conexões, essa ativação é propagada para outras unidades (MÜLLER; GUIDO, 2016; RASCHKA, 2015; ROJAS, 2013; SCHMID, 1994).

Redes Perceptron Multicamadas (*Multilayer Perceptron- MLP*) são as mais utilizadas e conhecidas. Neste tipo de rede, as unidades de processamento são dispostas verticalmente em diversas camadas. Há conexões entre as unidades das camadas adjacentes. A camada inferior é denominada camada de entrada (*input layer*), porque as ativações das unidades nessa camada representam a entrada da rede. A camada superior é denominada camada de saída (*output layer*). Quaisquer camadas entre as duas são denominadas camadas ocultas (*hidden layers*). A figura 24 mostra uma rede MLP com três camadas. Se a rede tiver um número de camadas superior a três, podemos denominá-la rede neural profunda (MÜLLER; GUIDO, 2016; RASCHKA, 2015; ROJAS, 2013; SCHMID, 1994).

Durante o processamento, as ativações são propagadas das unidades de entrada, através das unidades ocultas, até as unidades de saída. Para cada unidade j , as ativações multiplicadas pelos pesos $a_i w_{ij}$ são somadas e um parâmetro de *bias* O_j é adicionado. A rede aprende adaptando os pesos das conexões entre as unidades, até que a saída correta seja produzida. Um método muito utilizado é a retropropagação (*backpropagation*) que realiza uma busca gradiente descendente (*gradient descent*) na superfície do erro (MÜLLER; GUIDO, 2016; RASCHKA, 2015; ROJAS, 2013; SCHMID, 1994).

Figura 24 – Perceptron multicamadas com 3 camadas

Fonte: Raschka (2015)

Na figura 25 o gradiente descendente está representado como uma "descida da montanha" até que seja alcançado um mínimo, local ou global. A cada iteração, damos um passo na direção oposta ao gradiente, em que o tamanho do passo é determinado pela taxa de aprendizagem, assim como pela inclinação do gradiente (RASCHKA, 2015; SKANSI, 2018; MÜLLER; GUIDO, 2016; ROJAS, 2013; SCHMID, 1994).

Figura 25 – Gradiente descendente

Fonte: Raschka (2015)

Resumidamente, podemos afirmar que, durante o treinamento com o algoritmo de retropropagação, a rede opera em uma sequência de dois passos. Primeiro, um padrão é apresentado à camada de entrada da rede. A atividade resultante flui através da rede, camada por camada, até que a resposta seja produzida pela camada de saída. No segundo passo, a saída obtida é comparada à saída desejada para esse padrão particular. Se esta não estiver correta, o erro é calculado. O erro é retropropagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados conforme o erro é retropropagado (RASCHKA, 2015; SKANSI, 2018; MÜLLER; GUIDO, 2016; ROJAS, 2013; SCHMID, 1994).

A figura 26 ilustra o funcionamento do algoritmo de retropropagação em uma rede neural simples com três camadas. A camada de entrada tem dois neurônios: o neurônio cujo valor inicial é $x_A = 0.23$; e o neurônio B com valor inicial $x_B = 0.82$. A resposta esperada (*target*) para esse conjunto de dados de treinamento (x_A e x_B) é 1. Os pesos w_i estão especificados em cada ligação. Como é possível constatar, essa não é uma rede Perceptron, pois ambas as camadas, oculta e de saída, contêm uma função de ativação logística definida por:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (25)$$

O primeiro passo é calcular as saídas y_C e y_D . E depois usar ambas como entradas do neurônio F para calcular y_F tal como na equação 28 (SKANSI, 2018):

$$y_C = \sigma(0.23 * 0.1 * +0.82 * 0.4) = \sigma(0.351) = 0.5868 \quad (26)$$

$$y_D = \sigma(0.23 * 0.5 + 0.82 * 0.3) = \sigma(0.361) = 0.5892 \quad (27)$$

$$y_F = \sigma(0.5868 * 0.2 + 0.5892 * 0.6) = \sigma(0.4708) = 0.6155 \quad (28)$$

O passo seguinte é calcular a função de erro que, nesse caso, é o erro quadrático médio, como mostra a equação 30 (SKANSI, 2018):

$$E = \frac{1}{2}(t - y)^2 = \frac{1}{2}(1 - 0.6155)^2 = 0.0739 \quad (29)$$

Figura 26 – Algoritmo de retropropagação em uma rede neural de 3 camadas

Fonte: Skansi (2018)

Obtido o erro quadrático, calculamos as derivadas do erro E em relação aos pesos. Como o algoritmo de retropropagação segue na direção oposta ao passo para frente, é mais fácil começar o cálculo por w_5 . Queremos saber de que modo uma mudança em w_5 afeta E e desejamos considerar as mudanças que o minimizem. Para calcular a derivada parcial de E em relação a w_5 usamos a regra da cadeia tal como na equação 30 (SKANSI, 2018):

$$\frac{\partial E}{\partial w_5} = \frac{\partial E}{\partial y_F} * \frac{\partial y_F}{\partial z_F} * \frac{\partial z_F}{\partial w_5} \quad (30)$$

A equação 31 mostra o cálculo da derivada parcial de E em relação a y_F (SKANSI, 2018):

$$\frac{\partial E}{\partial y_F} = -(t - y_F) = -1(1 - 0.61555) = -0.3844 \quad (31)$$

Por sua vez, o cálculo da derivada de y_F em relação a z_F corresponde à equação 32 (SKANSI, 2018):

$$\frac{\partial y_F}{\partial z_F} = y_F(1 - y_F) = 0.6155(1 - 0.6155) = 0.2365 \quad (32)$$

O cálculo de z_F pode ser feito conforme equação 34 (SKANSI, 2018):

$$z_F = y_C * w_5 + y_D * w_6 \quad (33)$$

Como w_6 pode ser tratado como constante, temos que a derivada de z_F em relação a w_5 pode ser calculada conforme equação 34 (SKANSI, 2018):

$$\frac{\partial z_F}{\partial w_5} = y_C * 1 + y_D * 0 = 0.5868 \quad (34)$$

A partir dos três últimos resultados, calculamos a derivada parcial do erro E em relação ao peso w_5 conforme equação 35 (SKANSI, 2018):

$$\frac{\partial E}{\partial w_5} = -0.3844 * 0.2365 * 0.5868 = -0.0533 \quad (35)$$

Usando a regra geral de atualização de pesos e uma taxa de aprendizagem $\eta = 0.7$ obtemos o valor atualizado de w_5 mostrado na equação 36 (SKANSI, 2018):

$$w_5^{novo} = w_5^{velho} - \eta * \frac{\partial E}{\partial w_5} = 0.2 - (0.7 * 0.0533) = 0.2373 \quad (36)$$

Após calcularmos w_5 e w_6 prosseguimos para as camadas anteriores, atualizando os demais pesos. Depois que todos estiverem atualizados, damos um novo passo para frente com os novos pesos, de modo a verificar se o erro de fato diminuiu. Isso vale para cada amostra de dados do treinamento (SKANSI, 2018).

De forma simplificada, podemos dizer que uma Rede Neural Recorrente (*Recurrent Neural Network* - *RNN*) é um laço (*loop*) que reutiliza quantidades computadas durante a iteração prévia. RNNs são construídas por meio da adição de conexões recorrentes à camada oculta de uma rede *feedforward*.

Nesse tipo de rede, cada camada se conecta à próxima camada, porém não há caminho de volta. Todas as conexões têm, portanto, a mesma direção, saindo da camada de entrada rumo à camada de saída. Por razões históricas, estas redes são frequentemente confundidas com redes MLP; mas essa denominação é equivocada do ponto de vista técnico, dado que perceptrons são puramente lineares, não contêm funções de ativação do tipo logística, representada pela equação 25 (JURAFSKY; MARTIN, 2014).

As partes *a* e *c* da figura 27 ilustram a principal diferença entre as redes *feedforward* e RNNs: no primeiro caso as conexões não são circulares. A parte *c* corresponde exatamente à parte *a*, sendo que na primeira os neurônios individuais foram condensados em retângulos. As saídas de cada camada das redes *feedforward* são denotadas por I , O e H e por $H_1, H_2, \dots, H_N$ quando adicionamos as conexões recorrentes. Os pesos na rede *feedforward* são denotados por w_x (da camada de entrada para a camada oculta), e por w_0 (da camada oculta para a camada de saída) (SKANSI, 2018).

Figura 27 – Adição de conexões recorrentes a uma rede *feedforward* simples

Fonte: Skansi (2018)

A estrutura recursiva de uma RNN simples (*Simple Recursive Network- SRN*) pode ser representada pelas equações 37 e 38:

$$h(t) = f_h(w_h^T h(t-1) + w_x^T x(t)) \quad (37)$$

$$y(t) = f_o(w_o^T h(t)) \quad (38)$$

onde f representa uma função não-linear, por exemplo, a função logística (f_h é a função não-linear da camada oculta, e f_o da camada de saída), e $y(t)$ é a saída da rede no tempo t . Diferentemente do que ocorre nas redes *feedforward*, além de $y(t)$, calculamos também $h(t)$ que representa as entradas da conexão recorrente. Este tipo de SRN é denominado rede Elman (SKANSI, 2018).

SRNs têm um problema: embora deveriam teoricamente ser capazes de reter no tempo T informações sobre entradas vistas há muitos períodos, na prática, estas dependências de longo prazo são impossíveis de serem aprendidas. Isso ocorre devido ao problema do desaparecimento do gradiente (*vanishing gradient problem*), um efeito também observado em redes *feedforward* com muitas camadas de profundidade: à medida que adicionamos camadas a uma rede, ela se torna eventualmente impossível de treinar (CHOLLET, 2017).

As Redes de Memória de Longo Prazo (*Long Short Term Memory- LSTM*) foram concebidas para resolver o problema do desaparecimento do gradiente. A figura 28 compara os dois tipos de rede: LSTMs e SRNs. Uma diferença notável é que as SRNs possuem uma ligação $h(t)$ de uma unidade para outra, enquanto as LSTMs possuem o mesmo $h(t)$ e também $C(t)$, denominado memória de estado, que representa a memória de longo prazo

do modelo. Diferentes filtros decidem o que deve ser armazenado ou adicionado à memória de estado (SKANSI, 2018; CHOLLET, 2017).

Figura 28 – Unidades SRN e LSTM

Fonte: Skansi (2018)

RNNs podem processar uma grande quantidade de sequências etiquetadas, de modo a prever as etiquetas das próximas sequências, sendo portanto aplicáveis à etiquetagem morfossintática (SKANSI, 2018).

A utilidade da etiquetagem morfossintática deve-se à grande quantidade de informações que nos fornece sobre a palavra e seu contexto. Saber se uma palavra é substantivo ou verbo nos permite conhecer bastante sobre as prováveis palavras vizinhas (substantivos são precedidos por artigos e adjetivos, verbos são precedidos por substantivos) e sobre a estrutura sintática do contexto da palavra (substantivos são parte de sintagmas nominais), o que faz da etiquetagem morfossintática uma peça importante da análise sintática (*syntactic parsing*) (JURAFSKY; MARTIN, 2014).

4.3.3 Reconhecimento de entidades nomeadas

Conforme mostrado na figura 21, a etapa posterior à etiquetagem morfosintática dos *tokens* é o reconhecimento de entidades nomeadas (*Named-Entity Recognition - NER*) que abrange identificar cada menção à entidade e classificá-la por tipo. A definição dos tipos de entidades nomeadas depende da aplicação; normalmente, são pessoas, locais e organizações, mas também pode haver interesse em outras entidades, tais como genes e proteínas (JURAFSKY; MARTIN, 2014).

A técnica básica de NER é a segmentação do texto (*text chunking*), formando palavras e sequências etiquetadas, tal como na figura 29, na qual as caixas menores mostram os *tokens* e respectivas etiquetas e as maiores mostram os *chunks*. As etiquetas da figura pertencem ao conjunto de etiquetas do Penn Treebank Project (BIRD; KLEIN; LOPER, 2009).

Usando essa técnica, segmentamos o texto e etiquetamos as entidades que possam participar de relações entre si. Tipicamente identificamos sintagmas nominais (*Noun Phrases- NP*), compostos por nomes, nomes próprios ou ainda sequências de nomes próprios. Toda unidade cujo núcleo é um nome, ou seja, um substantivo, é denominada sintagma nominal (BIRD; KLEIN; LOPER, 2009).

Figura 29 – Segmentação e etiquetagem nos níveis dos *tokens* e dos *chunks*

Fonte: Bird, Klein e Loper (2009)

O primeiro passo é quebrar as sentenças, usando uma ou mais regras gramaticais. Podemos estabelecer, por exemplo, se nosso objetivo for identificar sintagmas nominais, que a gramática corresponda a uma sequência de nomes próprios com tamanho indefinido.

Vamos aplicá-la a uma sentença previamente etiquetada com classes gramaticais:

```
from nltk import RegexpParser
tagged_sent= [('O', 'ART'), ('Município', 'NPROP'), ('de', 'NPROP'),
('São', 'NPROP'), ('José', 'NPROP'), ('dos', 'NPROP'), ('Campos', 'NPROP'),
('recebeu', 'V'), ('R$', 'CUR'), ('300.000,00', 'NUM'), ('da', 'N'),
('União', 'NPROP'), ('Federal', 'NPROP'), ('para', 'PREP'), ('obras', 'N'),
('na', 'ADJ'), ('rede', 'N'), ('de', 'PREP'), ('esgotos', 'N')
# definição das regras gramaticais
gramatical = '\n'.join([
    'NE: {<NPROP>+}',
    'VAL: { <CUR><NUM>}'
])
# aplicação das regras gramaticais
from nltk import RegexpParser
analiseGramatical = nltk.RegexpParser(gramatical)
trees1 = analiseGramatical.parse(tagged_sent)
for tree in trees1:
    print(tree)
>>('O', 'ART')
(NE Município/NPROP de/NPROP São/NPROP José/NPROP dos/NPROP Campos/NPROP)
('recebeu', 'V')
(VAL R$/CUR 300.000,00/NUM)
('da', 'N')
(NE União/NPROP Federal/NPROP)
('para', 'PREP') ('obras', 'N')
('na', 'ADJ') ('rede', 'N')
('de', 'PREP') ('esgotos', 'N')
```

Nesse caso, foi possível identificar duas entidades nomeadas (NE): "Município de São José dos Campos" e "União Federal". À primeira vista parece-nos que a `gramatical` seja adequada à identificação das organizações envolvidas nas transferências de recursos para saneamento básico. Aproveitamos também para identificar valores repassados (VAL),

dado que o etiquetador CRF reconheceu corretamente R\$ como símbolo de moeda (CUR), e 300.000,00 como número (NUM).

Usando a mesma sentença etiquetada, é possível identificar não apenas as entidades nomeadas e valores transferidos, como também as relações (REL) entre elas. A seguir mostramos como aperfeiçoar a gramática¹, partindo da premissa de que as entidades mencionadas possam corresponder a mais de uma estrutura gramatical.

```
gramatica2 = '\n'.join([
    'NE: {<NPROP>+}',
    'NE: {<N>?<NPROP>+}',
    'VAL: { <CUR><NUM>}',
    'REL: {<NE> <.*>* <IN.*> <.*>* <NE>}'
])
from nltk import RegexpParser
analiseGramatica2 = nltk.RegexpParser(gramatica2)
trees2= analiseGramatica2.parse(tagged_sent)
print(trees2)
(S
  O/ART (NE
    Município/NPROP de/NPROP São/NPROP José/NPROP dos/NPROP Campos/NPROP)
    recebeu/V
    (VAL R\$/CUR 300.000,00/NUM)
  da/N
  (NE Unia~o/NPROP Federal/NPROP)
  para/PREP obras/N
  na/ADJ rede/N
  de/PREP esgotos/N)
```

A segmentação do texto em sintagmas nominais e o reconhecimento de entidades nomeadas podem ser feitos por meio de regras e expressões regulares, tal como demonstrado, mas o processo torna-se muito mais eficiente quando apoiado por algoritmos de aprendizado de máquina.

Os primeiros esquemas de segmentação de textos com aprendizado de máquina eram bem simples e identificavam apenas sintagmas nominais e verbais. Somente após

a CoNLL-2000 (*Conference on Natural Language Processing - CoNLL 2000*), é que um *corpus*, com um conjunto completo de tipos de *chunks*, foi introduzido. Desde então, muitas abordagens de aprendizado de máquina têm sido desenvolvidas. Dentre os modelos de melhor desempenho encontram-se as Máquinas de Vetores de Suporte e os Modelos Ocultos de Markov (BIRD; KLEIN; LOPER, 2009; JURAFSKY; MARTIN, 2014; SANG; BUCHHOLZ, 2000).

Embora a segmentação de textos tenha evoluído muito em outras línguas, o mesmo não ocorreu no português. Um dos poucos exemplos de *corpus* apropriado para segmentação de textos é o conjunto Bosque, que integra o projeto Floresta Sintática, um banco sintático (*treebank*) em português, brasileiro e europeu, e anotado automaticamente pelo analisador de textos PALAVRAS. Floresta Sintática é composto por diferentes conjuntos: Bosque, Selva, Amazônia e Floresta Virgem. O conjunto Bosque é o único revisado por linguistas, para corrigir os erros do analisador automático. Bosque é composto por dois *corpora*, cada um deles um subconjunto de *corpora* maiores, derivados de artigos de jornais. O maior *corpus* é o CETENFolha, composto por textos, extraídos do jornal "Folha de São Paulo" (AFONSO *et al.*, 2002).

Para os propósitos dessa pesquisa, as entidades nomeadas relevantes são os municípios beneficiados pelas transferências de recursos da União Federal. Também pode ser interessante identificar os órgãos responsáveis pelas transferências. E obviamente os montantes transferidos. No entanto, dada a estrutura muito peculiar das sentenças do DOU, não refletida em *corpora* cujos textos versem sobre temas de interesse amplo, utilizaremos regras gramaticais tipicamente associadas à presença de sintagmas nominais, em vez de algoritmos de aprendizado automático, para a segmentação de textos.

4.4 Identificação de padrões de escalamento

O segundo objetivo específico é identificar padrões de escalamento nas transferências de recursos da União Federal para os municípios, na modalidade saneamento básico, em relação ao tamanho dos municípios beneficiários.

Como vimos anteriormente, à medida que as cidades crescem em tamanho, observamos ganhos (ou economias) de escala e consequente decréscimo no volume ocupado pela infraestrutura urbana. Se consideramos a população $N(t)$, como medida do tamanho da

cidade no tempo t , os esperados ganhos de escala na infraestrutura das cidades podem ser expressos na forma de uma lei de potência do tipo $Y(t) = Y_0 N(t)^\beta$ onde $\beta \approx 0.8 < 1$, N denota volume de infraestrutura, e Y_0 é uma constante de normalização (BETTENCOURT *et al.*, 2007; BETTENCOURT, 2013).

Ganhos de escala na infraestrutura de saneamento básico significam um decréscimo nos custos, proporcional ao tamanho da cidade, levando a uma (esperada) relação de lei de potência entre os gastos com saneamento básico e o tamanho da cidade, medido em termos de sua população.

Para entender a importância de verificar se um fenômeno segue uma lei de potência, também denominada lei de escala, comecemos por recordar que muitas leis físicas são leis de potência, da qual a mais conhecida é a lei da gravitação de Newton, que é uma lei de potência em r , distância entre duas massas m_1 e m_2 : $F(r) = Gm_1m_2r^{-2}$. Outro exemplo de lei de potência é a difusão clássica, na qual a média quadrática dos deslocamentos de uma partícula escala com o tempo da forma $\langle (x(t) - x_0)^2 \rangle \approx t$.

Leis de potência são onipresentes nas ciências. Onde quer que haja uma relação de escalamento (*scaling relation*), ela é refletida pela existência de uma lei de potência, ou, de forma equivalente, funções de escalamento são sempre leis de potência (THURNER; HANEL; KLIMEK, 2018).

Um objeto é dito escalável se ele permanecer "invariante" sob transformações de escala do tipo $x \rightarrow \lambda x$. Frequentemente, esses objetos são representados por funções do tipo $f(x)$. Uma relação de escalamento estará presente se existirem funções contínuas e reais, f e g , tais que, sob uma transformação de escala do tipo $x \rightarrow \lambda x$, mantenha-se a relação (THURNER; HANEL; KLIMEK, 2018):

$$f(\lambda x) = g(\lambda)f(x) \quad (39)$$

onde λ é um número real positivo. A equação de escalamento tem uma solução real do tipo:

$$f(x) = b(g(x))^c \quad (40)$$

onde $g(x) = x^c$, b e c são constantes reais e g é denominada função de escalamento.

Objetos escaláveis aparecem em diferentes contextos. Eles podem assumir a forma de uma função, uma estrutura, uma lei física, ou uma função de distribuição que descreve a estatística de um sistema ou de um processo temporal. Estruturas que exibem escalamento são denominadas fractais. Elas podem ser fractais físicos, como um litoral ou uma rede

livre de escala, ou fractais matemáticos, como a curva de Koch ou o atrator de Lorentz. Além disso, diversas leis físicas são leis de potência. Elas também estão na descrição estatística de sistemas complexos estocásticos, nos quais o escalamento aparece nas funções de distribuição de quantidades observáveis de sistemas dinâmicos ou processos (THURNER; HANEL; KLIMEK, 2018).

No âmbito dessa pesquisa, a esperada lei de potência na infraestrutura de saneamento urbano baseia-se na premissa de que cidades sejam fractais físicos. A partir dela, queremos explicar a distribuição espacial dos serviços de saneamento básico no Brasil, em especial, a distribuição espacial do sistema de coleta de esgotos, muito concentrada na região Sudeste e nas áreas mais urbanizadas do País.

Dada a dificuldade de contabilizar todos os gastos com saneamento básico por município, pois há diversas fontes de recursos, que se superpõem em determinadas situações, vamos analisar somente as transferências da União Federal para os municípios, na modalidade saneamento básico. As fontes de dados foram detalhadas na subseção "Plano de Dados Numéricos e Textuais".

Sejam a e b os parâmetros que melhor ajustam uma lei de potência, ou seja, uma função $y = Ax^B$, onde $A = e^a$ e $B = b$, a um conjunto de dados (x_i, y_i) . Para calcular os valores de a e b , devemos minimizar o erro quadrático total expresso na equação 42:

$$E^2 = \sum_{i=1}^n \epsilon^2 = \sum_{i=1}^n [y_i - f(x_i)]^2 = \sum_{i=1}^n [y_i - Ax_i^B]^2 \quad (41)$$

Para tanto, devemos calcular as derivadas parciais do erro quadrático em relação a a e b e igualá-las a zero, tal como na equação 42, que equivale a calcular os parâmetros de uma regressão linear, expressos nas equações 43 e 44:

$$\frac{\partial E^2}{\partial a} = 0; \frac{\partial E^2}{\partial b} = 0 \quad (42)$$

$$b = \frac{n \sum_{i=1}^n (\ln x_i \ln y_i) - \sum_{i=1}^n (\ln x_i) \sum_{i=1}^n (\ln y_i)}{n \sum_{i=1}^n (\ln x_i)^2 - (\sum_{i=1}^n \ln x_i)^2} \quad (43)$$

$$a = \frac{\sum_{i=1}^n (\ln y_i) - b \sum_{i=1}^n (\ln x_i)}{n} \quad (44)$$

onde $A = e^a$ e $B = b$.

Para ajustar uma reta ao conjunto de dados (x_i, y_i) , onde x_i corresponde à população dos municípios, e y_i aos recursos transferidos pela União aos municípios, na modalidade saneamento básico, utilizaremos a biblioteca Numpy do Python.

Numpy possui uma função denominada `polyfit(x, y, deg)`, capaz de identificar o melhor ajuste de um polinômio de grau "deg" a um conjunto de pontos dados pelos argumentos x e y . A função `polyfit` retorna uma lista de coeficientes do polinômio ajustado, na qual o primeiro elemento é o coeficiente do termo de mais alto grau, e o último elemento é o termo constante (LANGTANGEN; LANGTANGEN, 2009).

Por exemplo, dados x_i e y_i , `polyfit(x, y, 1)` retorna os coeficientes, angular e linear, respectivamente, a e b , do polinômio de primeiro grau $y = ax + b$, que melhor se ajustem aos dados. A função `polyfit` também nos permite calcular a matriz de covariância dos coeficientes estimados. A diagonal dessa matriz contém as variâncias estimadas para cada coeficiente. Para o polinômio de grau 1, $y = ax + b$, a matriz de covariância será uma matriz quadrada de dimensão $(2, 2)$.

Para calcular os coeficientes estimados a e b e respectivos desvios-padrão, aplicamos a função `polyfit` aos logaritmos de x_m e y_m , que são valores de x e y a partir de um determinado patamar x_{min} :

```
import numpy as np
a, b = np.polyfit(np.log10(xm), np.log10(ym), 1)
V = np.cov(np.log10(xm), np.log10(ym))
std_0 = np.sqrt(V[0][0])
std_1 = np.sqrt(V[1][1])
print ("coef_angular = {:.2f} +/- {:.2f}".format(a, std_0))
print ("coef_linear = {:.2f} +/- {:.2f}".format(b, std_1))
```

Uma das formas de avaliar a qualidade do ajuste do modelo é por meio do coeficiente de determinação R^2 . O valor de R^2 varia entre 0 e 1, indicando, em termos percentuais, em que medida o modelo consegue explicar os valores observados. Para calculá-lo fazemos:

```
import numpy as np
R_squared = np.corrcoef(np.log10(xm), np.log10(ym))[0, 1]**2
print("R_quadrado_ = {:.2f}".format(R_squared))
```

5 Resultados obtidos

5.1 Softwares para extrair dados do DOU

5.1.1 Softwares baseados em regras gramaticais

Esse subitem pretende ser uma explicação sobre a estrutura e o funcionamento dos softwares para extrair dados do DOU no formato PDF, cujos códigos estão no Anexo A. Eles foram testados em três tipos de instrumentos jurídicos que regem as transferências de recursos da União Federal para os municípios: (i) extrato de convênio; (ii) extrato de contrato de repasse; e (iii) termo aditivo ao convênio.

As etapas de preparação do texto podem ser assim descritas. Em primeiro lugar, o arquivo previamente convertido em TXT é aberto, lido linha por linha, e dele são removidos caracteres não essenciais ao processamento posterior.

Em seguida, o texto é processado pelas funções "prepararPLN1" ou "prepararPLN2" que segmentam o texto em *tokens* e removem dados não essenciais: datas, e-mails, números de processo, endereços de sites, e lixo, definido como o conjunto de *tokens* com menos de dois caracteres. Para a quebra do texto em *tokens*, foi escolhida a função da NLTK `WhiteSpaceTokenizer()`, que considera apenas os espaços em branco entre eles. Os *tokens* passam então pelos dois etiquetadores morfossintáticos selecionados, CRF e NLPNET.

Para utilizar o etiquetador CRF precisamos carregar o modelo previamente treinado ("model.crf.tagger"):

```
crf_tagger= nltk.tag.CRFTagger()
crf_tagger.set_model_file('model.crf.tagger')
tagged = crf_tagger.tag(tokens)
```

Ao buscarmos etiquetadores baseados em redes neurais, identificamos a NLPNET, uma biblioteca Python baseada em duas redes: uma MLP e uma convolucional. A primeira é um etiquetador morfossintático e a segunda um anotador de papéis semânticos. Ambos os modelos foram treinados em uma versão do Mac-Morpho revisada pelos desenvolvedores (FONSECA; ROSA; ALUÍSIO, 2015).

Para utilizar o etiquetador NLPNET, cuja acurácia é de 93,66% nos dados de teste, é necessário carregar a biblioteca e o modelo para o português ("pos-pt"). Como o etiquetador é pré-treinado e aceita entradas em formato *string*, sua aplicação é simples e

direta. Ele quebra o texto em sentenças e também as transforma em *tokens*. No entanto, como o formato de saída é uma lista de listas de tuplas, não adequado às demais etapas de processamento de texto, é necessário convertê-lo para uma lista de tuplas:

```
import nlpnet
nlpnet.set_data_dir('pos-pt')
tagger = nlpnet.POSTagger()
tagged= tagger.tag(text)
from functools import reduce
tagged1= reduce(lambda x,y: x+y, tagged)
```

A etapa seguinte à etiquetagem é a segmentação do texto em sintagmas nominais, de modo a identificar as entidades nomeadas, constituídas tipicamente por nomes próprios ou sequências de nomes próprios. No *software* do etiquetador NLPNET, adotamos como premissa que o nome do município fosse constituído por até dois nomes próprios (NPROP) e normalmente precedida de um substantivo (N) seguido de preposição (PREP): "Município de" ou "Cidade de". Para obter os valores transferidos, supusemos que eles fossem expressos por até dois substantivos (N) seguidos de um número (NUM), porque o NLPNET classifica os dois caracteres referentes à moeda ("R" e "S") como substantivos distintos. Estas premissas foram traduzidas em regras gramaticais e expressões regulares:

```
gramatica = '\n'.join([
    'NE: {<N><PREP><NPROP>?<NPROP>}',
    'VAL: { <N><N>?<NUM>}'
])
from nltk import RegexpParser
parser = nltk.RegexpParser(gramatica3)
trees = parser.parse(sentences)
for tree in trees:
    print(tree)
```

No *software* baseado no etiquetador CRF, adotamos as mesmas regras gramaticais do caso anterior para extrair o nome do município. Para extrair os valores transferidos, supusemos que fossem expressos por um substantivo (N) seguido de um número (NUM), porque o etiquetador CRF reconhece corretamente os caracteres referentes à moeda ("R" e

"S") como um único substantivo. Estas premissas foram traduzidas em regras gramaticais e expressões regulares.

```
gramatica1 = '\n'.join ([
    'NE: {<N><PREP><NPROP>?<NPROP>}',
    'VAL: { <CUR><NUM>}',
])

from nltk import RegexpParser
parser = nltk.RegexpParser(gramatica1)
trees2 = parser.parse(sentences)
for tree in trees2:
    print(tree)
```

Nos dois *softwares*, há funções para extração dos municípios beneficiários e valores transferidos. Os resultados obtidos, por meio do etiquetador CRF associado à gramática1, na extração de entidades nomeadas e valores, tendo como fonte de dados um excerto do DOU referente a um Extrato de Termo Aditivo, foram:

```
def extrair_entidades_nomeadas(t):
    entidades_nomeadas = []
    if hasattr(t, 'label') and t.label:
        if t.label() == 'NE':
            entidades_nomeadas.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                entidades_nomeadas.extend
                    (extrair_entidades_nomeadas(child))
    return entidades_nomeadas

entidades_nomeadas=[]
for tree in parser.parse(sentences):
    entidades_nomeadas.extend(extrair_entidades_nomeadas(trees))
entidades_nomeadas= set((entidades_nomeadas))
print(entidades_nomeadas)
-> {'cidade de Tangará da Serra/MT,'}
```

```

def extrair_valores_transferidos(t):
    valores_transferidos= []
    if hasattr(t, 'label') and t.label:
        if t.label() == 'VAL':
            valores_transferidos.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                valores_transferidos.extend
                    (extrair_valores_transferidos(child))
    return valores_transferidos

valores_transferidos= []
for tree in parser.parse(sentences):
    valores_transferidos.extend(extrair_valores_transferidos(trees))
valores_transferidos = set((valores_transferidos))
print(valores_transferidos)
-> {'R\$ 8.542.710,03'}

```

Apesar dos resultados preliminares satisfatórios, a metodologia empregada para extração dos dados deixa a desejar em alguns aspectos comentados no próximo subitem. Para tornar a extração de dados mais eficiente e acurada, vamos desenvolver um modelo de rede LSTM bidirecional (BLSTM), a ser treinado como etiquetador morfossintático diretamente nos textos do DOU. Este mesmo modelo pode ser estendido ao reconhecimento de entidades nomeadas, com ou sem o uso de regras gramaticais ou expressões regulares.

5.1.2 Modelo BLSTM

Nesse subitem mostraremos como desenhar e treinar um modelo BLSTM, a ser utilizado na 'etiquetagem morfosintática, cujo código Python encontra-se no Anexo A.

As células simples LSTM processam o texto em um único sentido, normalmente da esquerda para a direita, mas as células BLSTM também levam em conta o sentido oposto, permitindo assim que o modelo revele mais padrões, pois há mais dados de entrada. Em outras palavras, o modelo BLSTM leva em conta não apenas a sequência de *tokens* após o *token* de interesse, mas também a sequência anterior.

Para implementar o modelo BLSTM, utilizamos Keras, um *framework* de alto nível que não requer conhecimentos sobre os algoritmos de funcionamento interno das redes neurais profundas. O pré-processamento dos dados, incorporados na camada de entrada, foi realizado com o auxílio de ferramentas da biblioteca Pandas do Python (MCKINNEY, 2011; GULLI; PAL, 2017).

Concluído o pré-processamento dos dados, passamos ao desenho do modelo. É possível escolher dentre dois tipos de Interface de Programação de Aplicações (*Application Program Interface* - *API*): sequencial e funcional. A API sequencial permite desenhar os modelos camada por camada, e funciona bem na maioria das situações, embora não permita construir modelos com camadas compartilhadas ou que tenham múltiplas entradas ou saídas. Em contraposição, a API funcional confere maior flexibilidade à construção de modelos com camadas que se conectem a quaisquer outras.

Usando uma API sequencial, desenhamos o modelo BLSTM, com quatro camadas, se excluirmos da contagem a camada de entrada, quais sejam: (i) incorporação de palavras (*word embeddings*); (ii) descarte (*dropout*); (iii) BLSTM; e (iv) camada densa de saída.

```
model= Sequential()
model.add(InputLayer(input_shape=(max_len, )))
model.add(Embedding(input_dim=n_words+1, output_dim= 128,
input_length = max_len))
model.add(Dropout(0.2))
model.add(Bidirectional(LSTM(units= 256, return_sequences=True,
recurrent_dropout=0.1)))
model.add(Dense(n_pos, activation= "softmax"))
```

O tensor da camada de entrada tem como uma de suas dimensões o tamanho máximo das sentenças, que depende de conhecimento prévio da distribuição de tamanhos das sentenças do texto a ser analisado.

Logo após a camada de entrada, há uma camada de incorporação de palavras que requer ao menos dois parâmetros: a quantidade de *tokens* ou palavras + 1; e a dimensionalidade da incorporação de palavras, que corresponde ao tamanho dos vetores densos (CHOLLET, 2017).

Para entender o que são vetores densos, recordemos que *embedding* funciona como um dicionário que mapeia índices inteiros (correspondentes a palavras específicas) em vetores densos. Ou seja, ele toma esses inteiros como entrada, busca-os em um dicionário interno e retorna os vetores associados. Diferentemente dos vetores obtidos por *one-hot encoding*, os vetores do *word embeddings* são de baixa dimensionalidade, compostos por números reais representados em ponto flutuante. Outra diferença entre eles é que os vetores densos da incorporação de palavras são aprendidos a partir dos dados (CHOLLET, 2017).

Por sua vez, a camada de descarte é um recurso para evitar sobreajuste (*overfitting*) do modelo aos dados de treinamento. Tipicamente, o erro nos dados de validação é muito menor quando se usa *dropout*, mas tendo como contrapartida o custo de modelos maiores e muito mais iterações do algoritmo de treinamento. A taxa de descarte é normalmente definida por um valor entre 0.2 e 0.5 (CHOLLET, 2017).

A camada BLSTM foi definida com 256 neurônios. Como outras camadas do Keras ela pode ser executada de duas formas: pode retornar toda a sequência de sucessivas saídas ou somente a última saída. Caso adotemos a primeira opção é necessário estabelecer que:

```
return_sequences = True
```

A função de ativação presente na camada densa, que também é a camada de saída, é uma versão da função logística para mais de duas classes, definida pela equação 45 (SKANSI, 2018).

$$\zeta(z_j) = \frac{e^{z_j}}{\sum_{n=1}^N e^{z_k}}, j = 1, \dots, N \quad (45)$$

Essa função, denominada *softmax*, transforma um vetor z , com valores reais arbitrários, em um vetor cujos valores situam-se entre 0 e 1, e que somados totalizam 1. O motivo para sua utilização na última camada de uma rede neural profunda, para classificação em múltiplas classes, é que as saídas podem ser interpretadas como as probabilidades de cada classe (SKANSI, 2018).

Antes do treinamento é necessário compilar o modelo. A compilação requer três parâmetros: (i) função de erro (*loss function*) - define como a rede medirá seu desempenho, e como fará para se orientar na direção certa; (ii) otimizador - mecanismo pelo qual a rede se atualiza com base nos dados e na sua função de erro; e (iii) métrica - para monitorar os resultados durante a etapa de treinamento e teste (SKANSI, 2018).

Entropia cruzada categórica é a função de erro normalmente utilizada em problemas de classificação em várias classes. Ela minimiza a distância entre as distribuições de probabilidade, a produzida pela rede e a dos *targets* (CHOLLET, 2017).

Na escolha do otimizador, podemos selecionar um dentre os algoritmos de aprendizagem adaptativa mais avançados: AdaGrad (Duchi, Hazan e Singer (2011)), RMSprop (Tieleman e Hinton (2012)), e Adam (Kingma e Ba (2014)).

AdaGrad é um algoritmo que ajusta a taxa de aprendizagem de forma proporcional à frequência dos parâmetros. Ela é maior para os parâmetros nos quais os gradientes anteriores eram pequenos (e sua soma também) e menor para os parâmetros nos quais os gradientes eram relativamente grandes. Dessa forma, a taxa de aprendizagem dos parâmetros mais frequentes é atualizada mais vezes, com valores menores; para os parâmetros menos frequentes, a taxa de aprendizagem é atualizada com valores e intervalos maiores. Essa característica o torna ideal para trabalhar com dados esparsos.

Adam é visto frequentemente como uma versão generalizada do AdaGrad (AdaGrad seria equivalente ao Adam mais a escolha de certos parâmetros). RMSprop também é similar ao Adagrad, mas substitui a soma dos quadrados dos gradientes anteriores, para atualização das taxas de aprendizagem adaptativas, por uma média móvel dos quadrados dos gradientes anteriores (RUDER, 2016).

RMSprop é um método adaptado ao treinamento em pequenos lotes de dados (*mini batches*) e por isso foi escolhido para o treinamento do modelo BLSTM. Usando RMSprop, entropia categórica cruzada, e acurácia como métrica de avaliação de desempenho do modelo, a compilação assume a seguinte forma:

```
model.compile(loss='categorical_crossentropy', optimizer="rmsprop",
              metrics=['accuracy'])
```

Usualmente, modelos de rede neural profunda não processam todo o conjunto de dados de treinamento de uma só vez; em vez disso, os dados são divididos em pequenos lotes.

As vantagens são: maior rapidez no treinamento e economia de memória computacional. Cada iteração sobre todo o conjunto de dados de treinamento é denominada época (*epoch*).

No código a seguir a rede começará a iterar no conjunto de dados de treinamento em lotes de 32 amostras, 80 vezes cada:

```
history = model.fit(X_train, np.array(y_train), batch_size= 32, epochs=80,
validation_split=0.1, verbose=1)
```

Um aspecto chave do treinamento dos modelos BLSTM aplicados a dados textuais é a incorporação de palavras. Há duas formas de implementá-la: (i) aprender os vetores densos enquanto realizamos a tarefa de nosso interesse, no caso a etiquetagem morfofossintática. Nesta forma começamos com vetores aleatórios e depois aprendemos os vetores densos, do mesmo jeito que aprendemos os pesos em uma rede neural; e (ii) incorporar ao nosso modelo *word embeddings* pré-treinadas em outra tarefa de aprendizado de máquina.

Dadas as peculiaridades do vocabulário do DOU, optamos por treinar os vetores densos usando um banco de dados especialmente criado para essa tarefa, no qual cada *token* foi classificado segundo a metodologia descrita no Manual do Mac-Morpho. Trata-se de um banco com 55 sentenças do DOU, a ser posteriormente ampliado de modo a incorporar toda a diversidade de textos relativos a transferências de recursos.

Antes do treinamento é necessário converter o banco de dados, originalmente em formato CSV, em um quadro de dados da biblioteca Pandas. O quadro 3 mostra suas 10 últimas linhas.

Quadro 3 – Sentenças do DOU com anotações morfofossintáticas

Sentenças	Tokens	Etiqueta
Sentença:55	Conveniente	N
Sentença:55	:	:
Sentença:55	FIDEL	NPROP
Sentença:55	CARLOS	NPROP
Sentença:55	SOUZA	NPROP
Sentença:55	DANTAS	NPROP
Sentença:55	CPF	N
Sentença:55	no	N
Sentença:55	811.548.105-00	NUM
Sentença:55	.	.

Fonte: Ludmila Deute Ribeiro, 2018

Apesar do banco de dados ser bastante limitado em quantidade de sentenças e *tokens*, o modelo BLSTM alcançou acurácia superior a 98%, nos dados de teste, e superior

a 97%, nos dados de validação. Ao utilizarmos o modelo treinado para etiquetar um texto desconhecido pela rede, mas com vocabulário e estrutura similares aos utilizados no treinamento, obtivemos o resultado mostrado no quadro 4.

Para estender o modelo BLSTM ao reconhecimento de entidades nomeadas podemos adotar as seguintes estratégias: (i) incluir, no banco de dados, uma coluna adicional de etiquetas do tipo IOB que indiquem a presença e o tipo das entidades nomeadas; ou (ii) associar os resultados obtidos pelo modelo BLSTM, aplicado à etiquetagem morfosintática, ao uso de expressões regulares e regras gramaticais (JURAFSKY; MARTIN, 2014).

Como exemplo do item (ii), se utilizássemos as regras gramaticais dos softwares anteriormente desenvolvidos para extração de dados, obteríamos o nome do município:

NE: :{<N><PREP><NPROP><NPROP>}

>>Município de Água Branca

Sobre as etiquetas IOB mencionadas no item (i), nesse esquema cada *token* dentro de um bloco de palavras(*chunk*) é marcado com três etiquetas: I (*inside*), O (*outside*), e B (*begin*). Um *token* é identificado como B se ele marca o início de uma entidade nomeada. Os tokens subsequentes dentro do bloco são identificados com etiquetas I. Todos os demais recebem etiquetas O.

Quadro 4 – Etiquetas previstas pelo modelo BLSTM

Tokens	Etiqueta Prevista
ESPÉCIE	NPROP
Contrato	NPROP
de	NPROP
Repasse	NPROP
no	N
868247/2018	N
,	,
firmado	PCP
pelo	PREP +
Município	N
de	PREP
Água	NPROP
Branca	NPROP
,	,
CNPJ	N
12.350/153/0001-48	NUM
;	;
junto	PREP
à	NPROP
União	NPROP
Federal	NPROP
,	,
por	PREP
intermédio	NPROP
do	NPROP
Ministério	NPROP
das	NPROP
Cidades	NPROP
,	,
representada	PCP
pela	PREP+
Caixa	NPROP
Econômica	NPROP
Federal	NPROP

Fonte: Ludmila Deute Ribeiro, 2018

5.1.3 Avaliação de acurácia dos modelos

Os *softwares* para extração dos dados do DOU, desenvolvidos no âmbito dessa pesquisa, deixam a desejar em alguns aspectos.

Em primeiro lugar, apesar da acurácia dos etiquetadores CRF e NLPNET ser superior a 90% nos dados de teste do Mac-Morpho, há pouca semelhança entre os *corpora*, DOU e Mac-Morpho. No DOU há uma grande quantidade de palavras em caixa alta ou iniciadas por letras maiúsculas, características que induzem os etiquetadores a classificá-las incorretamente como nomes próprios. Do que se conclui que, para contemplar essas especificidades, o treinamento dos etiquetadores deve ser realizado no próprio DOU.

Além disso, o reconhecimento de entidades nomeadas, com base em regras gramaticais, revelou-se um procedimento casuístico, não generalizável para todas as situações. Trata-se de uma solução artesanal, a ser adaptada caso a caso, o que impossibilita a avaliação de acurácia do modelo como um todo.

Em compensação, o modelo BLSTM alcançou acurácia superior a 98%, nos dados de teste, e a 97%, nos dados de validação, apesar do banco de dados utilizado para treinamento inicial ser bastante limitado em quantidade de *tokens* e pouco representativo do universo de sentenças do DOU.

A versão do modelo BLSTM apresentada nessa pesquisa é aplicável exclusivamente à etiquetagem morfossintática, mas é possível estendê-la ao reconhecimento de entidades nomeadas, associando-o ao uso de regras gramaticais e expressões regulares ou, alternativamente, ampliando o banco de dados, de modo a incluir uma coluna de etiquetas IOB que indiquem a presença e o tipo das entidades nomeadas associadas aos *tokens*.

Para aumentar a acurácia do modelo BLSTM, aplicado ao reconhecimento de entidades nomeadas, uma estratégia possível é implementar o modelo concatenado CRF e BLSTM, proposto por Jurafsky e Martin (2014) e mostrado na figura 30. A camada superior é uma camada CRF, cuja saída é um conjunto de etiquetas do tipo IOB que indicam não apenas início, meio e fim, mas também a presença de tipos específicos de entidades nomeadas, tais como pessoas (PER) e localidades (LOC).

Figura 30 – Modelo concatenado CRF-BLSTM com camada *embeddings* para NER

Fonte: Jurafsky e Martin (2014)

5.2 Padrões de escalamento observados nos dados do SIOP

As figuras 31 a 34 mostram os padrões de escalamento observados nos dados extraídos do SIOP, os quais se referem exclusivamente aos recursos não onerosos, classificados como: programados ("dotação inicial") e empenhados ("empenhado").

Para os valores programados, β estimado foi de aproximadamente: (i) (0.63 ± 0.54) para municípios com população superior a 2.000 habitantes; (ii) (0.92 ± 0.42) para municípios acima de 20.000 habitantes; e (iii) (1.18 ± 0.38) para municípios acima de 50.000 habitantes.

Para os valores empenhados, β estimado foi de aproximadamente: (i) (0.43 ± 0.57) para municípios com população superior a 2.000 habitantes; (ii) (0.57 ± 0.46) para municípios com mais de 20.000 habitantes; e (iii) (0.64 ± 0.41) para municípios com população superior a 50.000 habitantes.

Para municípios com população superior a 2.000 habitantes, os valores de R^2 que indicam, em termos percentuais, em que medida o modelo consegue explicar os valores

observados, foram de 0.38 para os valores programados ("dotação inicial"), e de 0.23 para os valores empenhados ("empenhado").

Para aqueles com população superior a 20.000 habitantes, os valores de R^2 foram de 0.33 para os valores programados ("dotação inicial"), e de 0.32 para os valores empenhados ("empenhado").

Para os municípios com mais de 50.000 habitantes, os valores de R^2 foram de 0.32 para os valores programados ("dotação inicial"), e de apenas 0.22 para os valores empenhados ("empenhado")

Em todos os casos, esse foi o melhor ajuste possível, obtido pela função polyfit do Python. Apesar do pequeno poder explicativo do modelo, consideramos que seja possível afirmar, com ressalvas, que os valores estimados de β mostram redução das transferências da União Federal para saneamento básico, proporcional ao tamanho dos municípios, em termos populacionais.

Os achados da pesquisa corroboram a hipótese de que ganhos de escala, proporcionais ao tamanho dos municípios, estejam sendo, de forma intencional ou não, aproveitados pela União Federal ao transferir recursos não onerosos para os municípios. Ou seja, os municípios mais populosos (e mais ricos) não têm sido aquinhoados com montantes superiores às suas reais necessidades, se considerarmos somente os recursos não onerosos. Se estivessem recebendo mais recursos do que o necessário, em detrimento dos municípios menores e mais pobres, essa seria uma possível explicação, ainda que parcial, para a concentração espacial dos serviços de saneamento básico.

Analisando apenas os municípios com mais de 50.000 habitantes, observamos, com base nos valores de β , que no momento do planejamento orçamentário, expresso pela dotação inicial, os municípios mais populosos (e mais ricos) têm sido aquinhoados com montantes superiores às suas necessidades. No entanto, no momento de efetivo comprometimento de recursos, expresso pelos valores empenhados, as economias de escala se impõem nas transferências direcionadas a esse mesmo grupo de municípios.

Resta-nos então analisar os recursos onerosos, para tentar explicar o porquê da distribuição espacial dos serviços de saneamento básico, em especial, da coleta de esgotos, ser tão concentrada na Região Sudeste e em áreas mais urbanizadas. Como vimos, há indícios de concentração dos recursos do FGTS na Região Sudeste.

Figura 31 – Transferências programadas pela União Federal para saneamento básico: β estimado para municípios com mais de 20.000 habitantes

Fonte: Ludmila Deute Ribeiro, 2018

Figura 32 – Transferências programadas pela União Federal para saneamento básico: β estimado para municípios com mais de 50.000 habitantes

Fonte: Ludmila Deute Ribeiro, 2018

Figura 33 – Valores empenhados pela União Federal para saneamento básico: β estimado para municípios com mais de 20.000 habitantes

Fonte: Ludmila Deute Ribeiro, 2018

Figura 34 – Valores empenhados pela União Federal para saneamento básico: β estimado para municípios com mais de 50.000 habitantes

Fonte: Ludmila Deute Ribeiro, 2018

6 Conclusões

Que dados podem ser usados para detectar padrões de escalamento nas transferências, da União Federal para municípios, na modalidade saneamento básico? Quais as principais características desses dados? Com que frequência são divulgados? São dados georreferenciados?

Foram identificadas duas fontes de dados: DOU e SIOP.

O SIOP é o sistema informatizado que suporta os processos de planejamento e orçamento da União Federal. O acesso é livre e gratuito por meio do Painel de Orçamento. Os dados são estruturados e abrangem apenas as transferências de recursos não onerosos destinados ao saneamento básico.

Não é necessário utilizar ferramentas de SQL para extraí-los, pois a interface do sistema é amigável o suficiente para ser entendida por qualquer pessoa com noções básicas sobre orçamento público.

Os dados são atualizados, pois o SIOP é interligado aos demais bancos de dados da União Federal. É possível identificar a localização geográfica dos gastos para cada ação orçamentária, uma forma pouco acurada de geolocalização, mas que pode ser útil em diversas aplicações.

Por sua vez, o DOU é um periódico da União Federal para publicação de atos oficiais, divulgado em formato PDF, de 2017 a 2019, por meio do Portal da Imprensa Nacional. As edições posteriores a 30/11/2017 também estão disponíveis no formato HTML.

O DOU é composto por textos em português brasileiro, ou seja, são dados não estruturados. Atualmente os textos são divulgados somente em formato digital, com frequência mínima diária. Futuramente pretende-se que sejam gerados de forma automática, sempre que transferências de recursos forem realizadas pela União Federal. Os dados que podem ser extraídos do DOU referem-se às transferências de recursos não onerosos, da União Federal para os municípios, e aos empréstimos realizados com recursos do FGTS.

Dadas essas características, quais ferramentas de processamento de linguagem natural seriam adequadas à mineração dos dados? E quais algoritmos de aprendizado automático seriam adequados à detecção de padrões nessas transferências?

A melhor ferramenta de processamento de linguagem natural para extração de dados do DOU, em formato PDF, relativos a nomes de municípios e valores transferidos, parece ser uma rede BLSTM, treinada nos textos do próprio DOU, com anotações morfossintáticas e de entidades nomeadas.

Com efeito, redes BLSTM representam o estado da arte na etiquetagem morfossintática e reconhecimento de entidades nomeadas. No entanto, para que os resultados obtidos sejam significativamente mais acurados, quando comparados aos resultados obtidos por outros algoritmos, o treinamento deve ser realizado em textos semelhantes, devidamente anotados com as classes gramaticais das palavras e as entidades nomeadas.

Resultados preliminares do treinamento de uma rede BLSTM para etiquetagem morfossintática estão no Capítulo "Resultados obtidos". Também foram definidas as etapas necessárias para estendê-la ao reconhecimento de entidades nomeadas.

Adicionalmente foram desenvolvidos dois *softwares* para extração de informações do DOU, testados em três tipos de instrumentos jurídicos: extrato de convênio; extrato de contrato de repasse; e termo aditivo ao convênio.

Em ambos, as etapas de pré-processamento de texto são executadas por algoritmos pré-treinados em português, disponíveis na biblioteca NLTK do Python. Para a etiquetagem morfossintática foram selecionados, pelo critério de acurácia, os algoritmos CRF e NLPNET, ambos treinados no Mac-Morpho, um *corpus* em português brasileiro.

No entanto, a quebra do texto em segmentos, essencial ao reconhecimento das entidades nomeadas (nomes de municípios) e dos valores transferidos pela União Federal, foi prejudicada pela escassez de *corpora* em português brasileiro, com anotações de estrutura sintática, sem os quais torna-se inviável a utilização de algoritmos de aprendizado supervisionado para tal fim.

Sendo assim, e dado que, tanto o vocabulário quanto a estrutura dos textos do DOU são muito específicos, em nada semelhantes aos textos do Mac-Morpho, optamos pelo uso de regras gramaticais e expressões regulares para identificação dos nomes de municípios e dos valores transferidos.

Os dados, referentes ao período analisado, mostram quais padrões de escalamento das transferências em função do tamanho dos municípios?

Vimos que economias de escala na infraestrutura de saneamento básico implicam redução de custos, de forma proporcional ao tamanho das cidades, levando a uma (esperada) lei de escala que relaciona gastos com saneamento básico ao tamanho das cidades.

Para verificar a existência dessa lei de escala, utilizamos dados extraídos do SIOP, referentes ao período 2011-2018, os quais abrangem somente recursos não onerosos. Não foi possível desagregar os dados sobre saneamento básico por componente, ou seja, não foi possível analisar separadamente o esgotamento sanitário.

Os valores estimados de β para "dotação inicial" e "empenhado", categorias utilizadas na LOA, mostram redução das transferências da União Federal para saneamento básico, de forma proporcional ao tamanho dos municípios, em termos populacionais.

Ou seja, os resultados obtidos corroboram a hipótese de que ganhos de escala, proporcionais ao tamanho dos municípios, estejam sendo aproveitados, de forma intencional ou não, pela União Federal, ao transferir recursos não onerosos para os municípios, na modalidade saneamento básico. Os municípios mais populosos (e mais ricos) não têm sido aquinhoados com montantes superiores às suas reais necessidades.

No entanto, essas conclusões não são extensíveis aos recursos onerosos destinados ao saneamento básico. Na verdade, há indícios de concentração destes recursos, sobretudo dos recursos do FGTS, na Região Sudeste. Segundo fontes oficiais, houve, em 2016, uma grande concentração dos recursos do FGTS (68,5%) para atendimento de propostas da Região Sudeste. A causa da concentração regional teria sido a dificuldade de acesso ao crédito pelos municípios das demais regiões. Padrão semelhante foi observado em 2018, ano em que 60% dos recursos foram destinados à Região Sudeste.

Para analisar os padrões de escalamento dos empréstimos com recursos do FGTS, em relação ao tamanho dos municípios beneficiários, seria necessário extrair do DOU os dados correspondentes, desagregados por município, no período de 2011 a 2018. O principal obstáculo à execução dessa tarefa é a indisponibilidade de edições do DOU, fora do período 2017 a 2019. Além disso, o acesso em PDF é restrito a excertos de texto, classificados por tipo de instrumento jurídico, tal como mostrado na figura 19.

Não há informações disponíveis sobre a aplicação dos recursos do FAT no saneamento básico, desagregados por região, ou estado da federação.

De que forma os resultados obtidos podem contribuir para o aperfeiçoamento das políticas públicas de saneamento urbano?

Os resultados obtidos não corroboram a hipótese de que as transferências de recursos não onerosos, da União Federal para os municípios, estejam reforçando a concentração espacial da oferta de serviços de saneamento urbano na região Sudeste e nas áreas mais urbanizadas do País. Os municípios mais populosos (e mais ricos) não têm sido aquinhoados com montantes superiores às suas reais necessidades.

No entanto, há indícios de concentração de recursos onerosos, sobretudo dos recursos do FGTS, na Região Sudeste. O que nos leva a supor que os empréstimos às empresas de saneamento, públicas e privadas, estejam reforçando a concentração espacial da oferta de serviços de saneamento básico, sobretudo do esgotamento sanitário, na região Sudeste.

O próximo passo seria investigar os recursos onerosos, FGTS e FAT, de forma sistemática, usando a mesma metodologia usada para os recursos não onerosos. Mas esses indícios são suficientes para repensar de imediato o papel desempenhado pelo FGTS no financiamento do saneamento básico.

Que outras metodologias podem ser utilizadas para ampliar o controle social e aperfeiçoamento da gestão das políticas de saneamento urbano?

Analisando o marco legal do saneamento básico, chamou-nos a atenção o fato da legislação determinar que os usuários providenciem a conexão entre suas residências e as redes de serviços de abastecimento de água ou de coleta de esgotos, quando disponíveis, o que gera despesas adicionais, além do pagamento da tarifa, e requer por vezes soluções complexas de engenharia, nas áreas com urbanização esparsa ou ao contrário, muito densa e precária, caso das comunidades de baixa renda. Na ausência de redes públicas de saneamento básico, são admitidas soluções individuais de abastecimento de água e de afastamento e destinação final dos esgotos sanitários.

Ao projetarmos os possíveis impactos desses dispositivos legais, não nos deveria causar surpresa a oferta de esgotamento sanitário ser tão concentrada nas regiões mais ricas e de alta densidade populacional do território brasileiro. Porque, ao estabelecer essas normas, os formuladores de políticas incluíram os usuários finais no planejamento e no financiamento do sistema de saneamento básico, tornando-o um problema perverso, persistente e não completamente explicitado. A característica essencial dos problemas perversos é que eles não podem, na prática, ser resolvidos por um planejador central.

Para resolver problemas perversos nas cidades não bastam *big data* e engenharia de sistemas. São necessários também modelos de comportamento humano. E a previsão de cadeias detalhadas de comportamentos têm a propriedade fundamental de serem praticamente impossíveis em cidades grandes, qualquer que seja a quantidade de dados disponíveis. É provável que várias soluções para problemas perversos no ambiente urbano, baseadas em diferentes modelos de comportamento, sejam testadas sem que se chegue a uma solução global para os problemas.

Por exemplo, a instituição de multas para os usuários que não se conectarem às redes de esgoto, quando disponíveis, resolveria o problema do acesso ao saneamento básico no Brasil? Alguns diriam que sim, outros que não, dado ser perfeitamente possível formular modelos de comportamento humano baseados nos quais multas sejam inócuas ou pouco efetivas. Particularmente, em áreas urbanas muito densas e precárias, habitadas por comunidades de baixa renda, a conexão de uma habitação individual às redes de esgoto próximas pode ser muito dispendiosa, complicada, ou até mesmo impossível sem a adesão das habitações vizinhas e, nesse caso, é razoável supor que o efeito das multas individuais nesses locais seja modesto ou nulo.

Dessa forma, se o marco regulatório mantiver as normas que incluem os usuários finais no planejamento e financiamento do saneamento básico, talvez a abundância de dados apenas capacite pessoas e organizações a resolver os seus problemas. E o caminho lógico, a partir da aceitação dessa premissa, seria a descentralização da gestão da política de saneamento básico, apoiada pelo uso de grandes volumes de dados, de modo a subsidiar o processo decisório participativo no âmbito dos próprios municípios e bairros.

Referências¹

- AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá (c) tica: a treebank for portuguese. In: ELRA. *quot; In Manuel González Rodrigues; Carmen Paz Suarez Araujo (ed) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)(Las Palmas de Gran Canaria Espanha 29-31 de Maio de 2002) Paris: ELRA. [S.l.], 2002. Citado na página 109.*
- ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*, APS, v. 74, n. 1, p. 47, 2002. Citado na página 59.
- ANDRIS, C.; BETTENCOURT, L. M. Development, information and social connectivity in côte d'ivoire. *Infrastructure Complexity*, Springer International Publishing, v. 1, n. 1, p. 1, 2014. Citado 3 vezes nas páginas 68, 69 e 70.
- BARABÁSI, A.-L. Network science book. *Boston, MA: Center for Complex Network, Northeastern University. Available online at: <http://barabasi.com/networksciencebook>*, 2014. Citado 2 vezes nas páginas 56 e 57.
- BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *science*, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citado na página 52.
- BATTY, M. Cities as complex systems: scaling, interactions, networks, dynamics and urban morphologies. Citeseer, 2008. Citado 3 vezes nas páginas 43, 45 e 46.
- BATTY, M. Building a science of cities. *Cities*, Elsevier, v. 29, p. S9–S16, 2012. Citado 3 vezes nas páginas 47, 48 e 64.
- BATTY, M. *The new science of cities*. [S.l.]: Mit Press, 2013. Citado 7 vezes nas páginas 48, 51, 52, 59, 60, 61 e 62.
- BATTY, M.; AXHAUSEN, K. W.; GIANNOTTI, F.; POZDNOUKHOV, A.; BAZZANI, A.; WACHOWICZ, M.; OUZOUNIS, G.; PORTUGALI, Y. Smart cities of the future. *The European Physical Journal Special Topics*, v. 214, n. 1, p. 481–518, 2012. Citado na página 62.
- BATTY, M.; LONGLEY, P. A. *Fractal cities: a geometry of form and function*. [S.l.]: Academic press, 1994. Citado 3 vezes nas páginas 48, 51 e 53.
- BATTY, M.; THURSTAIN-GOODWIN, M.; HAKLAY, M.; LLOYD, D.; BOOTT, R.; HYMAN, A.; BATTY, S.; TOMALIN, C.; CADELL, C.; FALK, N. *et al.* Producing boundaries and statistics for town centres. The Stationary Office, 2002. Citado na página 60.
- BERTALANFFY, L. V. General system theory. *New York*, v. 41973, n. 1968, p. 40, 1968. Citado na página 43.
- BETTENCOURT, L. M. The origins of scaling in cities. *science*, American Association for the Advancement of Science, v. 340, n. 6139, p. 1438–1441, 2013. Citado 5 vezes nas páginas 40, 46, 47, 70 e 110.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

BETTENCOURT, L. M. The uses of big data in cities. *Big Data*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 2, n. 1, p. 12–22, 2014. Citado 5 vezes nas páginas 64, 65, 66, 67 e 70.

BETTENCOURT, L. M.; LOBO, J.; HELBING, D.; KÜHNERT, C.; WEST, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 104, n. 17, p. 7301–7306, 2007. Citado 2 vezes nas páginas 40 e 110.

BETTENCOURT, L. M.; LOBO, J.; STRUMSKY, D.; WEST, G. B. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS one*, Public Library of Science, v. 5, n. 11, p. e13541, 2010. Citado na página 54.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. [S.l.]: "O'Reilly Media, Inc.", 2009. Citado 6 vezes nas páginas 87, 89, 90, 91, 106 e 109.

BRAND, S. *Whole Earth Discipline: Why Dense Cities, Nuclear Power, Transgenic Crops, Restored Wildlands, and Geoengineering Are Necessary*. [S.l.]: Penguin, 2010. Citado na página 53.

BRANTS, T. Tnt: a statistical part-of-speech tagger. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the sixth conference on Applied natural language processing*. [S.l.], 2000. p. 224–231. Citado na página 98.

BRASIL. *Constituição da República Federativa do Brasil*. 1998. Senado Federal: Centro Gráfico 1988. Citado 2 vezes nas páginas 20 e 72.

BRASIL. *Lei no 11.445, de 5 de janeiro de 2007*. 2007. <http://www.planalto.gov.br/ccivil.03/_Ato2007-2010/2007/Lei/L11445.htm>. Estabelece diretrizes nacionais para o saneamento básico; altera as Leis nos 6.766, de 19 de dezembro de 1979, 8.036, de 11 de maio de 1990, 8.666, de 21 de junho de 1993, 8.987, de 13 de fevereiro de 1995; revoga a Lei no 6.528, de 11 de maio de 1978, e dá outras providências. Citado 2 vezes nas páginas 20 e 22.

BRASIL. *Decreto no. 7.217, de 21 de junho de 2010*. 2010. <http://www.planalto.gov.br/ccivil.03/_Ato2007-2010/2010/Decreto/D7217.htm>. Regulamenta a Lei no. 11.445, de 5 de janeiro de 2007, que estabelece diretrizes nacionais para o saneamento básico, e dá outras providências. Citado na página 20.

BRASIL. *Decreto no. 7.404, de 23 de dezembro de 2010*. 2010. <http://www.planalto.gov.br/ccivil.03/_Ato2007-2010/2010/Decreto/D7404.htm>. Citado na página 20.

BRASIL. *Política Nacional de Resíduos Sólidos*. 2010. <http://www.planalto.gov.br/ccivil.03/_ato2007-2010/2010/lei/112305.htm>. Institui a Política Nacional de Resíduos Sólidos; altera a Lei no. 9.605, de 12 de fevereiro de 1988; e dá outras providências. Citado na página 20.

BRASIL. *Plano Nacional de Saneamento Básico - PLAN SAB*. 2014. <https://www.cidades.gov.br/images/stories/ArquivosSNSA/PlanSaB/plansab_texto_editado_para_download.pdf>. Citado na página 21.

- BRASIL. *Panorama dos Planos Municipais de Saneamento Básico no Brasil*. 2017. <<https://www.cidades.gov.br/saneamento-cidades/panorama-dos-planos-de-saneamento-basico-no-brasil>>. Citado 2 vezes nas páginas 22 e 23.
- BRASIL. *DOU - Diário Oficial da União*. 2018. Citado 2 vezes nas páginas 76 e 77.
- BRASIL. *Gasto Público Saneamento Básico 2016*. 2018. Citado 3 vezes nas páginas 27, 28 e 82.
- BRASIL. *Manual Técnico de Orçamento 2018*. 2018. <http://www.planejamento.gov.br/assuntos/orcamento-1/informacoes-orcamentarias/arquivos/MTOs/mto_atual.pdf/view>. Citado 4 vezes nas páginas 79, 80, 81 e 82.
- BRASIL. *Medida Provisória no 844 de 2018*. 2018. <<https://www.congressonacional.leg.br/materias/medidas-provisorias/-/mpv/133867>>. Citado na página 23.
- BRASIL. *Medida Provisória no 868 de 2018*. 2018. <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/Mpv/mpv868.htm>. Citado na página 24.
- BRASIL. *Relatório de Gestão de 2017 : Fundo de Amparo ao Trabalhador - FAT*. 2018. <<http://portalfat.mte.gov.br/execucao-financeira-do-fat/relatorio-de-gestao-do-fat/>>. Citado na página 27.
- BRASIL. *Relatório de Gestão do Exercício de 2017 do FGTS*. 2018. <http://www.caixa.gov.br/Downloads/fgts-relatorio-gestao/RELATORIO_GESTAO_FGTS_2017.pdf>. Citado 3 vezes nas páginas 27, 29 e 30.
- BRASIL. *SIOP - Sistema Integrado de Planejamento e Orçamento do Governo Federal*. 2018. Citado na página 85.
- CHAUHAN, T.; ALUVALU, R. Using big data analytics for developing crime predictive model. 2016. Citado 2 vezes nas páginas 67 e 68.
- CHOLLET, F. *Deep learning with python*. [S.l.]: Manning Publications Co., 2017. Citado 4 vezes nas páginas 104, 105, 118 e 119.
- CHRISTALLER, W. *Central places in southern Germany*. [S.l.]: Prentice-Hall, 1966. Citado 2 vezes nas páginas 48 e 69.
- CORRALO, G. da S. *Município: autonomia na federação brasileira*. [S.l.]: Jurua Editora, 2006. Citado na página 71.
- CUNHA, A. dos S. *Saneamento básico no Brasil: Desenho institucional e desafios federativos*. Brasília, 2011. Disponível em: <<http://hdl.handle.net/10419/91100>>. Citado 2 vezes nas páginas 21 e 24.
- CUNHA, D. F. D. Região metropolitana: Apenas uma estrutura territorial legalizada? 2015. Citado 2 vezes nas páginas 72 e 73.
- DUCHI, J.; HAZAN, E.; SINGER, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, v. 12, n. Jul, p. 2121–2159, 2011. Citado na página 119.

- ENAP, E. N. de A. P. *Regras Específicas do Ministério das Cidades*. 2015. [⟨http://repositorio.enap.gov.br/handle/1/3022⟩](http://repositorio.enap.gov.br/handle/1/3022). Citado 2 vezes nas páginas 31 e 32.
- ENAP, E. N. de A. P. *Regras Gerais para acesso aos recursos de saneamento*. 2015. [⟨http://repositorio.enap.gov.br/handle/1/1878⟩](http://repositorio.enap.gov.br/handle/1/1878). Citado na página 31.
- ERDOS, P.; RÉNYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, v. 5, n. 1, p. 17–60, 1960. Citado na página 55.
- FONSECA, E. R.; ROSA, J. L. G.; ALUÍSIO, S. M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, Springer, v. 21, n. 1, p. 2, 2015. Citado 2 vezes nas páginas 88 e 113.
- FORRESTER, J. W. Urban dynamics. *IMR; Industrial Management Review (pre-1986)*, Massachusetts Institute of Technology, Cambridge, MA, v. 11, n. 3, p. 67, 1970. Citado 3 vezes nas páginas 43, 44 e 45.
- FUJITA, M. *Krugman and AJ Venables (1999), The spatial economy: cities, region and international trade*. [S.l.]: MIT press: Cambridge MA, 2001. Citado na página 69.
- GABAIX, X. Zipf’s law for cities: an explanation. *The Quarterly journal of economics*, MIT Press, v. 114, n. 3, p. 739–767, 1999. Citado na página 52.
- GUARE, J. *Six degrees of separation: A play*. [S.l.]: Vintage, 1990. Citado na página 56.
- GULLI, A.; PAL, S. *Deep Learning with Keras*. [S.l.]: Packt Publishing Ltd, 2017. Citado na página 117.
- HARDENIYA, N. *NLTK essentials*. [S.l.]: Packt Publishing Ltd, 2015. Citado 3 vezes nas páginas 87, 88 e 98.
- JACKSON, M. O. *Social and economic networks*. [S.l.]: Princeton university press, 2010. Citado na página 54.
- JURAFSKY, D.; MARTIN, J. H. *Speech and language processing*. [S.l.]: Pearson London, 2014. v. 3. Citado 11 vezes nas páginas 93, 94, 95, 96, 103, 105, 106, 109, 121, 123 e 124.
- KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Citado na página 119.
- LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. Citado 2 vezes nas páginas 98 e 99.
- LANGTANGEN, H. P.; LANGTANGEN, H. P. *A primer on scientific programming with Python*. [S.l.]: Springer, 2009. v. 2. Citado na página 112.
- LOSCH, A. *et al.* Economics of location. Yale University Press, 1954. Citado 2 vezes nas páginas 48 e 69.
- MARGUTI, B. O.; COSTA, M. A. Governança metropolitana no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2016. Citado na página 73.

MARGUTI, B. O. O.; COSTA, M. A. O.; FAVARÃO, C. B. O. Brasil metropolitano em foco: desafios à implementação do estatuto da metrópole. Instituto de Pesquisa Econômica Aplicada (Ipea), 2018. Citado 3 vezes nas páginas 72, 73 e 74.

MCKINNEY, W. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, v. 14, 2011. Citado na página 117.

MILLER, J. H.; PAGE, S. E. Complex adaptive systems. *An Introduction to Computational Models of Social Life*, Princeton, NJ, 2007. Citado na página 46.

MORI, T.; NISHIKIMI, K.; SMITH, T. E. The number-average size rule: A new empirical relationship between industrial location and city size. *Journal of Regional Science*, Wiley Online Library, v. 48, n. 1, p. 165–211, 2008. Citado na página 69.

MOURA, R.; OLIVEIRA, S.; PÊGO, B. Escalas da urbanização brasileira. Instituto de Pesquisa Econômica Aplicada (Ipea), 2018. Citado 4 vezes nas páginas 35, 38, 39 e 40.

MÜLLER, A.; GUIDO, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. [S.l.]: O'Reilly Media, Inc., Sebastopol, 2016. Citado 3 vezes nas páginas 99, 100 e 101.

NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. *The structure and dynamics of networks*. [S.l.]: Princeton University Press, 2011. Citado na página 46.

NEWMAN, M. E. Random graphs as models of networks. *arXiv preprint cond-mat/0202208*, 2002. Citado na página 55.

NEWMAN, M. E. The structure and function of complex networks. *SIAM review*, SIAM, v. 45, n. 2, p. 167–256, 2003. Citado na página 54.

NUAIMI, E. A.; NEYADI, H. A.; MOHAMED, N.; AL-JAROODI, J. Applications of big data to smart cities. *Journal of Internet Services and Applications*, Springer, v. 6, n. 1, p. 25, 2015. Citado 2 vezes nas páginas 63 e 64.

PERKINS, J. *Python 3 text processing with NLTK 3 cookbook*. [S.l.]: Packt Publishing Ltd, 2014. Citado 2 vezes nas páginas 88 e 92.

RASCHKA, S. *Python machine learning*. [S.l.]: Packt Publishing Ltd, 2015. Citado 3 vezes nas páginas 99, 100 e 101.

RITTEL, H.; WEBBER, M. M. 2.3 planning problems are wicked. *Polity*, v. 4, p. 155–169, 1973. Citado na página 64.

ROJAS, R. *Neural networks: a systematic introduction*. [S.l.]: Springer Science & Business Media, 2013. Citado 3 vezes nas páginas 99, 100 e 101.

RUDER, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. Citado na página 119.

SANEAMENTO-SNIS, S. N. D. I. S. *Diagnóstico do Manejo de Resíduos Sólidos Urbanos*. [S.l.]: Ministério das Cidades/SNSA/PMSS Brasília, 2017. Citado 3 vezes nas páginas 33, 34 e 38.

SANEAMENTO-SNIS, S. N. D. I. S. *Diagnóstico dos Serviços de Água e Esgoto*. [S.l.]: Ministério do Desenvolvimento Regional/SNSA/PMSS Brasília, 2017. Citado 7 vezes nas páginas 24, 25, 26, 33, 34, 36 e 37.

SANG, E. F. T. K.; BUCHHOLZ, S. Introduction to the conll-2000 shared task: Chunking. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*. [S.l.], 2000. p. 127–132. Citado na página 109.

SCHMID, H. Part-of-speech tagging with neural networks. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 15th conference on Computational linguistics-Volume 1*. [S.l.], 1994. p. 172–176. Citado 3 vezes nas páginas 99, 100 e 101.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. [S.l.]: Elsevier Brasil, 2017. Citado na página 86.

SIOPDOC - Documentação Online do SIOP. 2018. <<https://www1.siop.planejamento.gov.br/siopdoc/doku.php/ead-siop:i-introducao>>. Accessed: 2019-01-28. Citado na página 78.

SKANSI, S. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*. [S.l.]: Springer, 2018. Citado 8 vezes nas páginas 100, 101, 102, 103, 104, 105, 118 e 119.

STERMAN, J. D. J. D. *Business dynamics: systems thinking and modeling for a complex world*. [S.l.: s.n.], 2000. Citado na página 44.

SUI, D. Z. Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers*, Wiley Online Library, v. 94, n. 2, p. 269–277, 2004. Citado na página 53.

THURNER, S.; HANEL, R.; KLIMEK, P. *Introduction to the theory of complex systems*. [S.l.]: Oxford University Press, 2018. Citado 2 vezes nas páginas 110 e 111.

TIELEMAN, T.; HINTON, G. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *COURSERA Neural Networks Mach. Learn*, 2012. Citado na página 119.

TOBLER, W.; MIELKE, H.; DETWYLER, T. Geobotanical distance between new zealand and neighboring islands. *BioScience*, American Institute of Biological Sciences Circulation, AIBS, 1313 Dolley Madison Blvd., Suite 402, McLean, VA 22101. USA, v. 20, n. 9, p. 537–542, 1970. Citado na página 53.

USP, C. I. para Computação Linguística da. *Conjunto de Etiquetas(Tagset) do Mac-Morpho, versão 10*. 2003. <<http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf>>. Citado na página 90.

WHITE, R.; ENGELN, G.; ULJEE, I. *Modeling cities and regions as complex systems: From theory to planning applications*. [S.l.]: MIT Press, 2015. Citado 3 vezes nas páginas 49, 50 e 51.

WIENER, N. Cybernetics. *Scientific American*, JSTOR, v. 179, n. 5, p. 14–19, 1948. Citado na página 43.

Apêndice A – Softwares para extração de informações do DOU

Software em Python com Etiquetador CRF

```
# importa bibliotecas necessárias
import nltk
import re

# treinamento do etiquetador CRF no corpus Mac-Morpho
from nltk.corpus import mac_morpho
tagged_sentences = mac_morpho.tagged_sents()
print (tagged_sentences[0])
print ("Tagged sentences: ", len(tagged_sentences))
print ("Tagged words:", len(nltk.corpus.mac_morpho.tagged_words()))
size = int(len(tagged_sentences) * 0.75)
train_sents = tagged_sentences[:size]
test_sents= tagged_sentences[size:]

crf_tagger= nltk.tag.CRFTagger()
crf_tagger.train(train_sents, 'model.crf_tagger')
crf_tagger.evaluate(test_sents)

#abertura do arquivo de texto
file = open('arquivo.txt', 'rt')
text = file.read()
for line in text:
    line= line.rstrip()
file.close()
texto = text.replace(";", "").replace("-", "").replace(".", "").replace(":", "").replace("|", "")

# função de pré-processamento e etiquetagem morfosintática
def preparar_PLN1(texto):
    tokenizer= nltk.tokenize.WhitespaceTokenizer()
    tokens = tokenizer.tokenize(texto)
    datas = [w for w in tokens if re.search(r'\d{1,2}[/-]\d{1,2}[/-]\d{2,4}', w)]
    emails = [w for w in tokens if re.search(r'@[A-Za-z0-9_]+', w)]
    sites = [ w for w in tokens if re.search( r"(\http://[^\s]+)", w)]
    processos= [w for w in tokens if re.search(r'(\?|\$|B-)?\d+(\?:[.:,\d+/\d+\d+\d+)', w)]
    tokens = [w for w in tokens if w not in emails and w not in sites
               and w not in datas and w not in processos]
    garbage = [ w for w in tokens if len(w)< 2]
    tokens = [w for w in tokens if w not in garbage]
    crf_tagger= nltk.tag.CRFTagger()
    crf_tagger.set_model_file('model.crf_tagger')
    tagged = crf_tagger.tag(tokens)
    return tagged
```

```

sentences = preparar_PLN1(text)
print(sentences)

#gramática para quebra de textos em sintagmas nominais e extração de valores
gramatical = '\n'.join ([
    'NE: {<N><PREP><NPROP>?<NPROP>}',
    'VAL: { <CUR><NUM>}',
])

from nltk import RegexpParser
parser = nltk.RegexpParser(gramatical)
trees = parser.parse(sentences)
for tree in trees:
    print(tree)

# extração de entidades nomeadas
def extrair_entidades_nomeadas(t):
    entidades_nomeadas = []
    if hasattr(t, 'label') and t.label:
        if t.label() == 'NE':
            entidades_nomeadas.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                entidades_nomeadas.extend(extrair_entidades_nomeadas(child))
    return entidades_nomeadas

entidades_nomeadas= []
for tree in parser.parse(sentences):
    entidades_nomeadas.extend(extrair_entidades_nomeadas(trees))
entidades_nomeadas= set((entidades_nomeadas))
print(entidades_nomeadas)

# extração de valores
def extrair_valores_transferidos(t):
    valores_transferidos= []
    if hasattr(t, 'label') and t.label:
        if t.label() == 'VAL':
            valores_transferidos.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                valores_transferidos.extend(extrair_valores_transferidos(child))
    return valores_transferidos

valores_transferidos= []
for tree in parser.parse(sentences):
    valores_transferidos.extend(extrair_valores_transferidos(trees))
valores_transferidos = set((valores_transferidos))
print(valores_transferidos)

```

Software em Python com Etiquetador NLPNET

```
# importa bibliotecas necessárias
import nltk
import re

# abre o arquivo de texto
file = open('arquivo.txt', 'rt')
text = file.read()
for line in text:
    line= line.rstrip()
file.close()
texto = text.replace(";", "").replace("-", "").replace("...", "").replace(":", "").replace(" |", "")

# função de pré-processamento e etiquetagem morfossintática
def preparar_PLN2(texto):
    tokenizer= nltk.tokenize.WhitespaceTokenizer()
    tokens = tokenizer.tokenize(texto)
    datas = [w for w in tokens if re.search(r'\d{1,2}[/-]\d{1,2}[/-]\d{2,4}', w)]
    emails = [w for w in tokens if re.search(r'@[A-Za-z0-9_]+', w)]
    sites = [ w for w in tokens if re.search( r"(\http://[\s]+)", w)]
    processos= [w for w in tokens if re.search(r'(\?\$:\B-)?\d+(?:[.:,\]\d+/\d+\d+\d+)', w)]
    tokens = [w for w in tokens if w not in emails and w not in sites
               and w not in datas and w not in processos]
    garbage = [ w for w in tokens if len(w)< 2]
    tokens = [w for w in tokens if w not in garbage]
    import nlpnet
    nlpnet.set_data_dir('pos-pt')
    tagger = nlpnet.POSTagger()
    tagged= tagger.tag(texto)
    from functools import reduce
    tagged1= reduce(lambda x,y: x+y, tagged)
    return tagged1
sentences= preparar_PLN2(text)
print(sentences)

gramatica2 = '\n'.join([
    'NE: {<N><PREP><NPROP>?<NPROP>}',
    'VAL: { <N><N>?<NUM>}'
])

from nltk import RegexpParser
parser = nltk.RegexpParser(gramatica2)
trees = parser.parse(sentences)
for tree in trees:
    print(tree)
```

```

# extração de entidades mencionadas
def extrair_entidades_nomeadas(t):
    entidades_nomeadas = []
    if hasattr(t, 'label') and t.label:
        if t.label() == 'NE':
            entidades_nomeadas.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                entidades_nomeadas.extend(extrair_entidades_nomeadas(child))
    return entidades_nomeadas

entidades_nomeadas= []
for tree in parser.parse(sentences):
    entidades_nomeadas.extend(extrair_entidades_nomeadas(trees))
entidades_nomeadas= set((entidades_nomeadas))
print(entidades_nomeadas)

# extração de valores
def extrair_valores_transferidos(t):
    valores_transferidos= []
    if hasattr(t, 'label') and t.label:
        if t.label() == 'VAL':
            valores_transferidos.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                valores_transferidos.extend(extrair_valores_transferidos(child))
    return valores_transferidos

valores_transferidos= []
for tree in parser.parse(sentences):
    valores_transferidos.extend(extrair_valores_transferidos(trees))
valores_transferidos = set((valores_transferidos))
print(valores_transferidos)

```

Apêndice B – Ajuste de uma lei de escala aos dados SIOP

```
# importa bibliotecas necessárias
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# importa dados SIOP e converte para DataFrame da biblioteca Pandas
sanea = pd.read_csv("Cidades.csv", sep= ";")
sanea = sanea.sort_values(by= "População", ascending = True)
sanea.head(10)

# localiza e retira as linhas com zero de dotação inicial
sanea.drop(sanea[sanea["Dotação Inicial"]==0].index, inplace=True)

# simplifica notação
x=sanea["População"]
y=sanea["Dotação Inicial"]

# Gráfico 1 - Gastos Programados com Infraestrutura de Saneamento por Tamanho da Cidade
sns.set_style( "whitegrid")
plt.figure(figsize= (10,6))
plt.scatter(x, y)
plt.xlabel("População dos Municípios")
plt.ylabel("Transferências Programadas da União Federal para os Municípios (R\ $ X 10^5)")

# calcula a média anual
x = x
y = y/7

Gráfico 2 - Gastos Programados com Infraestrutura de Saneamento por Tamanho da Cidade - LOG LOG
fig=plt.figure(figsize=(10,6))
plt.scatter(np.log10(x),np.log10(y))
plt.xlabel("log(População dos Municípios)", fontsize= 12)
plt.ylabel("log(Transferências Programadas da União para Municípios(R\ $ 10^5))", fontsize= 12)

#exclui cidades com população inferior a 20.000 habitantes
xmin=20000
maiores=sanea.drop(sanea[sanea["População"]<=xmin].index)
xm=maiores["População"]
ym=maiores["Dotação Inicial"]

#calcula coeficientes angular e linear e matriz de covariância e  $R^2$  para xmin=20.000
m, b = np.polyfit(np.log10(xm), np.log10(ym), 1)
```



```

V =np.cov(np.log10(xm), np.log10(ym))
std_0 = np.sqrt(V[0][0])
std_1= np.sqrt(V[1][1])
R_squared =np.corrcoef(np.log10(xm), np.log10(ym))[0, 1]**2
print ("beta_1 = {:.2f} +/- {:.2f}".format(m, std_0))
print ( "coef_linear_1= {:.2f} +/- {:.2f}".format(b, std_1))
print("R_quadrado_1 = {:.2f}".format(R_squared))
plt.plot(np.log10(xm), m*np.log10(xm)+b, 'k')

#plota xmin como reta vertical
plt.plot([np.log10(xmin),np.log10(xmin)], [4,8], '--k')

fig.text(0.4, 0.8, '$Pop_{min}$ = '+str(xmin)+'', beta = '+str(round(m,2))', fontsize= 12)
#exclui cidades com população inferior a 2.000 habitantes
xmin=2000
maiores=sanea.drop(sanea[sanea["População"]<=xmin].index)
xm=maiores["População"]
ym=maiores["Dotação Inicial"]

#calcula coeficiente angular, matriz de covariância e R^2 para xmin=2.000
m, b = np.polyfit(np.log10(xm), np.log10(ym), 1)
V =np.cov(np.log10(xm), np.log10(ym))
std_0 = np.sqrt(V[0][0])
std_1= np.sqrt(V[1][1])
R_squared =np.corrcoef(np.log10(xm), np.log10(ym))[0, 1]**2
print ("beta_0: {:.2f} +/- {:.2f}".format(m, std_0))
print ("coef_linear_0: {:.2f} +/- {:.2f}".format(b, std_1))
print("R_quadrado_0 = {:.2f}".format(R_squared))
plt.plot(np.log10(xm), m*np.log10(xm)+b, 'r')

#plota xmin como reta vertical
plt.plot([np.log10(xmin),np.log10(xmin)], [4,8.1], '--r')

fig.text(0.2, 0.83, '$Pop_{min}$ = '+str(xmin)+'', beta = '+str(round(m, 2))',
fontsize= 12,color='red')
# fig.text(0.4, 0.8, '$Pop_{min}$ = '+str(xmin)+' $\\alpha$ = '+str(round(m,2))'
\normalsize

```

Anexo A – Treinamento de Etiquetadores Morfossintáticos

```
# etiquetador Naive Bayes
tagged_sents = mac_morpho.tagged_sents()
size = int(len(tagged_sents) * 0.75)
train_data = tagged_sents[:size]
test_data= tagged_sents[size:]
from nltk.classify import NaiveBayesClassifier
from nltk.tag.sequential import ClassifierBasedPOSTagger
nbt= ClassifierBasedPOSTagger(train= train_data,
classifier_builder= NaiveBayesClassifier.train)
nbt.evaluate(test_data)
0.8355477755432383

# treinamento do etiquetador Unigrama
tagged_sents= nltk.corpus.mac_morpho.tagged_sents()
texto='O Município de Salvador recebeu recursos para saneamento básico'
tokens= nltk.word_tokenize (texto, language= 'portuguese')
unigram_tagger = nltk.tag.UnigramTagger(tagged_sents)
unigram_tagger.tag(tokens)
[(('O', 'ART'), ('Município', 'NPROPN'), ('de', 'PREP'), ('Salvador', 'NPROPN'),
('recebeu', 'V'), ('recursos', 'N'), ('para', 'PREP'),
('saneamento', 'N'), ('básico', 'ADJ')]]

# combinação (com backoff) dos etiquetadores, Unigrama, Bigrama, e Trigrama
tagged_sents = mac_morpho.tagged_sents()
size = int(len(tagged_sents) * 0.75)
train_data = tagged_sents[:size]
test_data= tagged_sents[size:]
t0 = nltk.DefaultTagger('N')
t1 = nltk.UnigramTagger(train_data, backoff=t0)
t2 = nltk.BigramTagger(train_data, backoff=t1)
t3 = nltk.TrigramTagger(train_data, backoff=t2)
# grava o modelo
from pickle import dump
output = open('mac_morpho.pkl', 'wb')
dump(t3, output, -1)
output.close()
# importa o modelo
from pickle import load
input = open('mac_morpho.pkl', 'rb')
tagger_backoff = load(input)
input.close()
print(tagger_backoff.evaluate(test_data))
0.8997277676950998
```

```

# etiquetador HMM
tagged_sents = mac_morpho.tagged_sents()
size = int(len(tagged_sents) * 0.75)
train_data = tagged_sents[:size]
test_data= tagged_sents[size:]
from nltk.tag import hmm
trainer = hmm.HiddenMarkovModelTrainer()
hmm_tagger = trainer.train_supervised(train_data)
hmm_tagger.evaluate(test_data))
0.5151107323294256

# etiquetador TnT com N=100 e etiqueta padrão igual a substantivo (N)
tagged_sents = mac_morpho.tagged_sents()
size = int(len(tagged_sents) * 0.75)
train_data = tagged_sents[:size]
test_data= tagged_sents[size:]
from nltk.tag import tnt
from nltk.tag import DefaultTagger
unk= DefaultTagger('N')
tnt_tagger= tnt.TnT(unk=unk, Trained= True, N=100)
tnt_tagger.train(train_data)
tnt_tagger.evaluate(test_data)
0.885950605778192

# etiquetador CRF
tagged_sentences = mac_morpho.tagged_sents()
size = int(len(tagged_sents) * 0.75)
train_data = tagged_sents[:size]
test_data= tagged_sents[size:]
from nltk.tag import CRFTagger
crf_tagger= nltk.tag.CRFTagger()
crf_tagger.train(train_sents,'model.crf_tagger')
crf_tagger.evaluate(test_sents)
0.9249031245401481

# aplicação do etiquetador pré-treinado NLPNET
import nlpnet
nlpnet.set_data_dir('pos-pt')
tagger_nlp = nlpnet.POSTagger()
texto= "O Município de São José dos Campos recebeu R$ 300.000,00 da
União Federal para obras na rede de esgotos"
tagged_nlp= tagger_nlp.tag(texto)
print(tagged_nlp)
[[('O', 'ART'), ('Município', 'N'), ('de', 'PREP'), ('São', 'NPROP'),
('José', 'NPROP'), ('dos', 'NPROP'), ('Campos', 'NPROP'), ('recebeu', 'V'),
('R', 'N'), ('$', 'N'), ('300.000,00', 'N'), ('da', 'PREP+ART'),
('União', 'NPROP'), ('Federal', 'NPROP'), ('para', 'PREP'), ('obras', 'N'),

```

```

('na', 'PREP+ART'),('rede', 'N'), ('de', 'PREP'), ('esgotos', 'N')]]

# Treinamento de uma rede BLSTM para etiquetagem morfossintática

# importa bibliotecas necessárias
import nltk
import pandas as pd
import numpy as np
import keras

# abre o banco de dados CSV e converte em DataFrame Pandas
data = pd.read_csv("dataset_DOU.csv", sep= ";", encoding= "latin1")

# conta quantidade de palavras e de classes gramaticais
words = list(set(data["Word"].values))
words.append("FIM")
n_words = len(words); n_words
pos = list(set(data["POS"].values))
n_pos =len(pos); n_pos

# define classe ObterSentença
class ObterSentença(object):
    def __init__(self, data):
        self.n_sent = 1
        self.data = data
        self.empty = False
        agg_func = lambda s: [(w, p) for w, p in zip(s["Word"].values.tolist(),
                                                    s["POS"].values.tolist())]
        self.grouped = self.data.groupby("Sentence #").apply(agg_func)
        self.sentences = [s for s in self.grouped]
    def get_next(self):
        try:
            s = self.grouped["Sentence: {}".format(self.n_sent)]
            self.n_sent += 1
            return s
        except:
            return None

# obtém uma sentença dividida em tokens com respectivas classes gramaticais
getter = ObterSentença(data)
sent = getter.get_next(); print(sent)

# obtém todas as sentenças e faz um histograma
sentences = getter.sentences
import matplotlib.pyplot as plt
plt.style.use("ggplot")
plt.hist([len(s) for s in sentences], bins=50)
plt.show()

```

```

# define tamanho máximo da sentença e cria dicionários para palavras e classes gramaticais
max_len = XX
word2idx = {w: i for i, w in enumerate(words)}
pos2idx = {w: i for i, w in enumerate(pos)}

np.random.seed(1)
from keras.preprocessing.sequence import pad_sequences
# mapeia as sentenças em uma sequência de números e depois preenche a sequência
X = [[word2idx[w[0]] for w in s] for s in sentences]
X = pad_sequences(maxlen=max_len, sequences=X, padding="post",
value=n_words - 1)

# mapeia as classes gramaticais em uma sequência de números e depois preenche a sequência
y = [[pos2idx[w[1]] for w in s] for s in sentences]
y = pad_sequences(maxlen=max_len, sequences=y, padding="post",
value= n_pos -1)

# one-hot encoding
from keras.utils import to_categorical
y = [to_categorical(i, num_classes=n_pos) for i in y]

#divisão dos dados em treinamento e teste
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15)

# importa bibliotecas Keras
from keras.models import Sequential
from keras.layers import Dense, LSTM, InputLayer, Bidirectional,
Dropout, Embedding

# modelo sequencial com 5 camadas
model= Sequential()
model.add(InputLayer(input_shape=(max_len, )))
model.add(Embedding(input_dim=n_words +1, output_dim= 128, input_length=max_len))
model.add(Dropout(0.1))
model.add(Bidirectional(LSTM(units= 256, return_sequences=True, recurrent_dropout=0.1)))
model.add(Dense(n_pos, activation= "softmax"))
model.summary()

# compilação e treinamento do modelo
model.compile(loss='categorical_crossentropy',
              optimizer="rmsprop",
              metrics=['accuracy'])
history = model.fit(X_train, np.array(y_train), batch_size= 32, epochs=80,
validation_split= 0.1, verbose=1)

```