ANÁLISE CRÍTICA DA FONTE DOS DADOS UTILIZADA

Principais dificuldades enfrentadas no processo de obtenção dos dados

Avaliação de cada arquivo .csv:

- ESTADOS_BRASILEIROS: os dados sobre área territorial foram coletados da tabela “Área territorial - Brasil, Grandes Regiões, Unidades da Federação e Municípios” do IBGE, de 2024, e estavam bem formatados. Os dados relativos aos CARs foram coletados da tabela de Cadastro Ambiental Rural do Ministério da Gestão e da Inovação em Serviços Públicos, de 2024, e estavam bem formatados. Foi feita uma pesquisa para tentar encontrar tabelas com todos os cadastros detalhados por estado e ano, mas não foram encontradas. Isso se deu, possivelmente, por questões de sigilo, a fim de evitar a divulgação de endereço, de CNPJ ou CPF responsáveis pela área, etc. O site do Sistema de Cadastro Ambiental Rural (SICAR) possui documentos no formato .txt informando todos os cadastros por UF, mas são diferentes .txt para diferentes estados e cada arquivo é de vasta extensão, ou seja, difícil tratamento. Vale dizer que o registro de CARs é ineficiente e problemático, pois ele registra mais áreas rurais do que o valor possível por UF, representando um grave problema para a gestão de dados por parte do poder público. Portanto, para uma análise mais concisa acerca da área agropecuária de cada estado, foram utilizados os dados do Censo Agropecuário do IBGE de 2017, dados que se apresentaram muito mais confiáveis, mesmo que sejam de 2017.

- POPULACAO_ANO: dados provenientes das tabelas “Padrão de vida e distribuição de rendimentos” do IBGE, de 2024, e estavam bem formatados;

- GRAU_POBREZA_ANO: os dados foram coletados também das tabelas “Padrão de vida e distribuição de rendimentos” do IBGE, de 2024, e estavam bem formatados;

- GRAU_INSTRUCAO_ANO: dados provenientes das tabelas “Educação” do IBGE, de 2024, e estavam bem formatados. Todavia, houve a falta de dados referentes aos anos de 2020 e 2021, provavelmente devido à pandemia da COVID-19;

- CADASTRO_DE_EMPREGADORES: os dados foram coletados da tabela de Cadastro de Empregadores que tenham submetido trabalhadores a condições análogas à de escravo do Ministério do Trabalho, popularmente conhecida como “Lista Suja”, com última atualização em outubro de 2025, e houve problemas devido à má formatação de certos campos.
    - Empregador: embora esse atributo não tenha sido utilizado no Modelo Relacional, ele estava presente em seu respectivo arquivo .csv. Nele,

vale notar que os campos não estão padronizados, pois há a utilização do CNPJ do cadastro no respectivo nome do empregador (veja a Figura 1);

	ID_CE	Ano_d	Sigla_	Empregador	CNPJ_CPF
2	1	2024	SP	41.297.068 GILBERTO ELENO BATISTA DOS SANTOS	41.297.068/0001-61
3	2	2024	MT	48.937.720 ROBERTO DOS SANTOS	48.937.720/0001-04
4	3	2023	SP	50.964.355 FLAVIO DONIZETI DOS SANTOS	50.964.355/0001-79
5	4	2024	PE	53.162.923 EVAMBIVALDO FERREIRA GONCALVES	53.162.923/0001-0E
6	5	2024	PR	54.890.003 ALEXANDRE JUCELINO ZUKOVSKI	54.890.003/0001-77
7	6	2024	MS	57.299.132 AIRTON DE ARAUJO GOMES	57.299.132/0001-83
8	7	2024	PE	A2 ENGENHARIA LTDA	54.951.803/0001-5C
9	8	2024	ES	ABEL PIONA BERNABÉ	076.802.477-30
10	9	2024	PE	ACADEMIA ACTTIONOLINDA LTDA	53.093.685/0001-24
11	10	2020	AM	ADALCIMAR DE OLIVEIRA LIMA	153.980.052-00

Figura 1: Falta de padronização no atributo Empregador

- Cod_classe: dos 688 cadastros no arquivo .csv, 46 possuíam o atributo CNAE mal formatado: o padrão esperado é “XXXX-X/XX”, de acordo com o código CNAE, mas alguns atributos vieram formatados como data, conforme mostra a Figura 2. Entretanto, como havia a tabela em formato PDF disponibilizada no site do Ministério do Trabalho, foi possível realizar a correção manual dos valores. Para tanto, todos os atributos CNAE mal formatados tiveram seus valores substituídos por “NULL” e foram isolados nas primeiras linhas do arquivo .csv, como mostra a Figura 3, para que facilitasse a substituição manual.

Trabalhad	CNAE	Decisao_ac
2	03/01/4399	
2	03/05/2391	
2	4520-0/06	
3	07/07/3314	
1	4744-0/99	
9	0161-0/03	
20	4120-4/00	

Figura 2: Exemplo de má formatação em campos do atributo CNAE

A	B	C	D	E	F	G	H	I	J
id_ce	ano_da_a	sigla_uf	empregad	estabeleci	trabalhad	cnae	div_cnae	grp_cnae	
1	2024	SP	41.297.06	ROD. PREF	2	NULL	NULL	NULL	
2	2024	MT	48.937.72	MT 235, ZC	2	NULL	NULL	NULL	
4	2024	PE	53.162.92	AV. MANO	3	NULL	NULL	NULL	
34	2023	BA	ALEXANDF	RUA 24 FE	2	NULL	NULL	NULL	
37	2024	SC	ALTENHOF	COMUNID	7	NULL	NULL	NULL	
57	2023	RS	ANGELO L	AV.FREDEF	1	NULL	NULL	NULL	
87	2021	RJ	ASA BRAN	RUA JORN	2	NULL	NULL	NULL	
92	2023	BA	BARRA FO	FAZENDA S	5	NULL	NULL	NULL	
101	2025	MA	BRUNO RC	ZONA RUR	6	NULL	NULL	NULL	

Figura 3: Isolamento das tuplas cujo valor do atributo CNAE foi substituído por “NULL”

- DIVISAO_CNAE, GRUPO_CNAE e CLASSE_CNAE: os dados foram coletados da tabela “Estrutura detalhada da CNAE 2.0: Códigos e denominações” do IBGE e houve problemas quanto à organização das informações na tabela. A Figura 4 mostra um exemplo (de um total de 26 ocorrências) da presença de textos não relacionados às tuplas no interior do arquivo .csv, referentes às colunas da tabela.

44				02.20-9	Produção florestal - florestas nativas
45			02.3		Atividades de apoio à produção florestal
46				02.30-6	Atividades de apoio à produção florestal
47	2.2 - Estrutura detalhada da CNAE 2.0: Códigos e denominações				
48					(continuação)
49	Seção	Divisão	Grupo	Classe	Denominação
50		3			PESCA E AQÜICULTURA
51			03.1		Pesca
52				03.11-6	Pesca em água salgada
53				03.12-4	Pesca em água doce
54			03.2		Aqüicultura

Figura 4: Presença de textos no interior da estrutura da tabela