SEADE

Conferência Nacional dos Agentes Produtores e Usuários de Dados

Modernização da produção estatística

Carlos Torres Freire
Diretor do Centro de Ciência de Dados para Estatísticas Públicas (CCDEP / SEADE)

03 de Dezembro de 2025

SEADE
Fundação Sistema Estadual de Análise de Dados

GOVERNO DO ESTADO DE SÃO PAULO

AGENDA

1) **Pontos de partida**
    - Era digital e a produção estatística
    - Princípios de qualidade de processos e produtos estatísticos
2) **Experimentos recentes do CCDEP / SEADE:**
    - Áudio, texto, imagem, localização e IA Generativa
3) **Centro de Ciência de Dados para Estatísticas Públicas (CCCDEP)**

Era digital e desafios na produção estatística

TRADICIONALMENTE

- Estatísticas via coleta primária de dados – censos, *surveys* e registros administrativos

NA ERA DIGITAL

- Dados em **grande volume, alta frequência e diversidade**
- **Novos métodos e tecnologias** para coletar, armazenar, tratar e analisar dados para produção de estatísticas

REGISTROS GERADOS DIGITALMENTE POR AÇÕES HUMANAS E INTERAÇÕES SOCIAIS E CAPTURADOS AUTOMATICAMENTE

REGISTROS ADMINISTRATIVOS DIGITAIS

Cadastros sociais: educação, saúde e assistência social

Informações fiscais: pagamentos, notas de transações

Documentos: Declarações de nascimentos e óbitos

SUBPRODUTOS DE AÇÕES EM EQUIPAMENTOS

Uso de celular

Transações financeiras

Bilhetagem em transporte público

Visitas a websites

Publicação de notícias e em redes sociais

COLETA AUTOMATIZADA POR SENSORES

Presença física: tráfego de veículos e pessoas ou iluminação

Remotos: satélites e veículos aéreos

Princípios de qualidade de processos e produtos estatísticos

PRODUTOS ESTATÍSTICOS

- **Relevância**: importância para o usuário, demanda da sociedade
- **Precisão**: descrição correta do fenômeno, validade
- **Atualidade**: distância da data de referência
- **Coerência**: comparabilidade, convenções, padrão conceitual, universo
- **Acessibilidade**: disponibilidade e transparência
- **Interpretabilidade**: informações adicionais, como metadados, metodologias, classificações

PROCESSOS ESTATÍSTICOS

- **Metodologia sólida**: conceitos, classificações, convenções nacionais e internacionais
- **Adequação** de instrumentos, técnicas e sistemas tecnológicos
- **Solicitação não excessiva** em coletas e uso de registros administrativos
- **Relação entre custo e eficiência**

Monitor Covid-19

Pesquisa diária sobre temas relacionados à pandemia

- Acompanhamento de sintomas; distanciamento social; atividade econômica e consumo; trabalho e renda; educação; vacina; expectativas da população.
- De 23/Março/2020 a Setembro/2021.
- **Coleta de dados por telefone**, utilizando Unidade de Resposta Audível (**URA**).
- **800 entrevistas diárias**, para constituir um painel móvel de 4.800 entrevistas
- **População do Estado de São Paulo**, com amostra representativa por local de residência, **sexo e idade**
- Desagregação por Estado, RMSP, município SP, demais municípios da RMSP e Interior; e **17 Departamentos Regionais de Saúde.**

Experiência levou a pesquisas de percepção:

- **Cultura**
- **Atividade física**
- **Turismo**
- **Endividamento**
- **Segurança Alimentar**
- **Cuidados no Domicílio**
- **Hábitos de Consumo por App e Internet**

Coleta de informações por robô via telefone e aprendizado de máquina para classificar textos

Análise de respostas abertas em formato de áudio a partir de pesquisas de percepção por telefone com a população

TRANSCRIÇÃO DE ÁUDIO	PREPARAÇÃO DE MENSAGENS	CLASSIFICAÇÃO DE TEXTO	VALIDAÇÃO DE TÓPICOS
Processo automatizado que converte mensagens de áudio em texto.	Adequação dos textos substituindo palavras equivalentes e removendo termos comuns.	Uso de modelo não supervisionado para classificar mensagens em tópicos principais.	Análise de textos outliers e definição de rótulos para cada tópico pela equipe do SEADE.

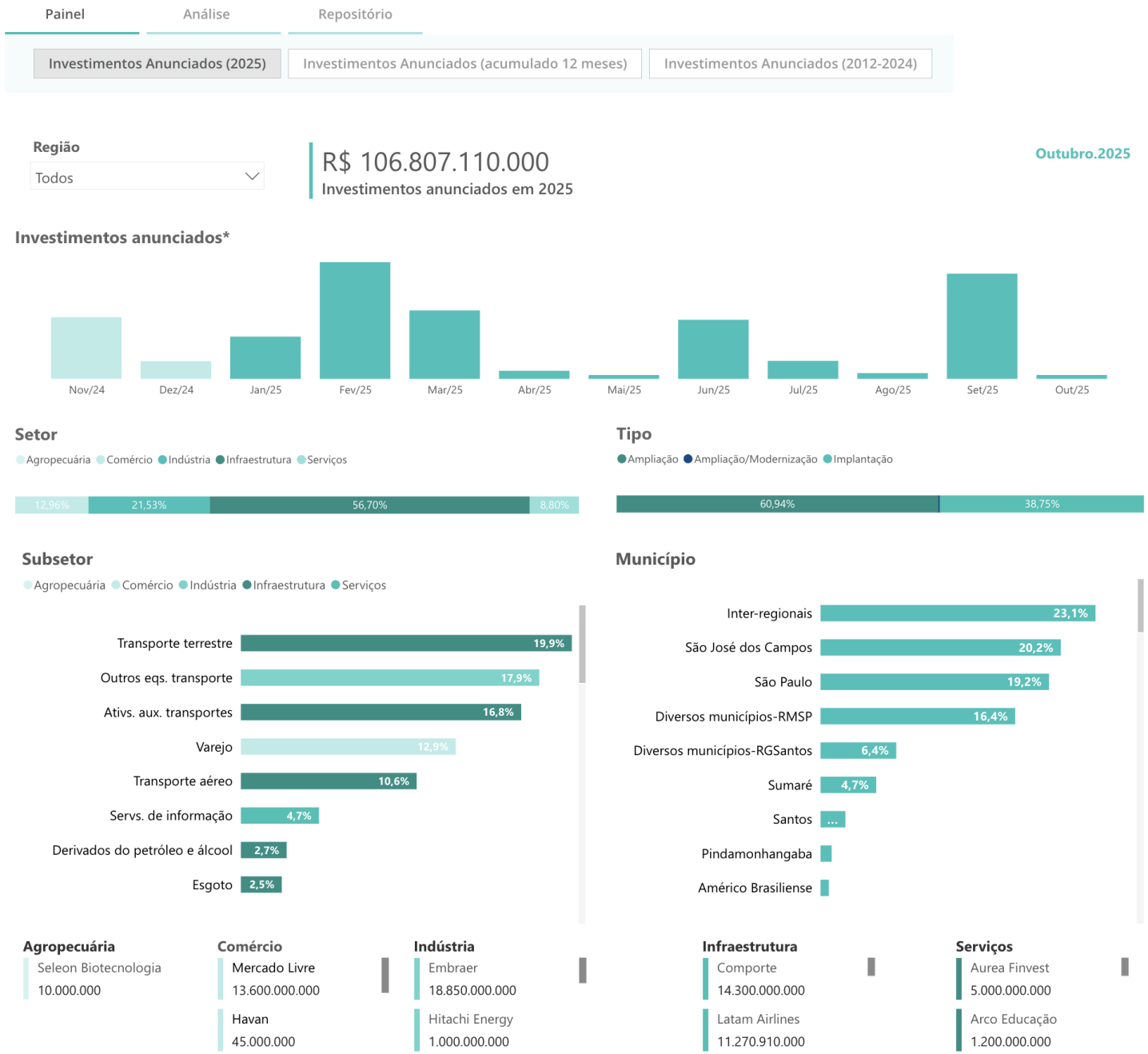
ATUALIDADE | RELEVÂNCIA | CUSTO / BENEFÍCIO

COERÊNCIA e COMPARABILIDADE

Exemplo internacional: Statistics Canada utiliza modelos de processamento de linguagem natural para classificar respostas sobre ocupações

SEADE INVESTIMENTOS

- Monitoramento sistemático e automatizado dos anúncios de investimentos de empresas privadas e públicas no Estado
- Captação diária de notícias de anúncios na imprensa
- Equipe de coleta confirma com empresas e completa informações como tipo, período, valor e local do investimento.



Raspagem de dados para monitorar investimentos

***Web scraping* de anúncios de investimentos em veículos de imprensa, com sistema multiagentes, e classificação de texto para triagem de notícias, com aprendizado de máquina, e produção de indicadores**

CAPTAÇÃO DE NOTÍCIAS

Monitoramento automatizado de anúncios na imprensa sobre investimentos

TRIAGEM POR IA

Modelos de aprendizado de máquina classificam textos sobre investimentos

PROCESSAMENTO DE INDICADORES

Organização dos dados por tipo, período, valor e localização do investimento

VALIDAÇÃO HUMANA

Equipe confirma informações com empresas e completa dados faltantes

Exemplo internacional: Projeto ESSnet Big Data do Eurostat classifica setorialmente e por ocupação anúncios de emprego em plataformas

ATUALIDADE RELEVÂNCIA

PRECISÃO COMPARABILIDADE

Leitura de imagens de declarações do Registro Civil

Modelos de aprendizado de máquina para leitura automatizada de imagens de declarações de nascimento, óbito e casamentos enviadas por cartórios

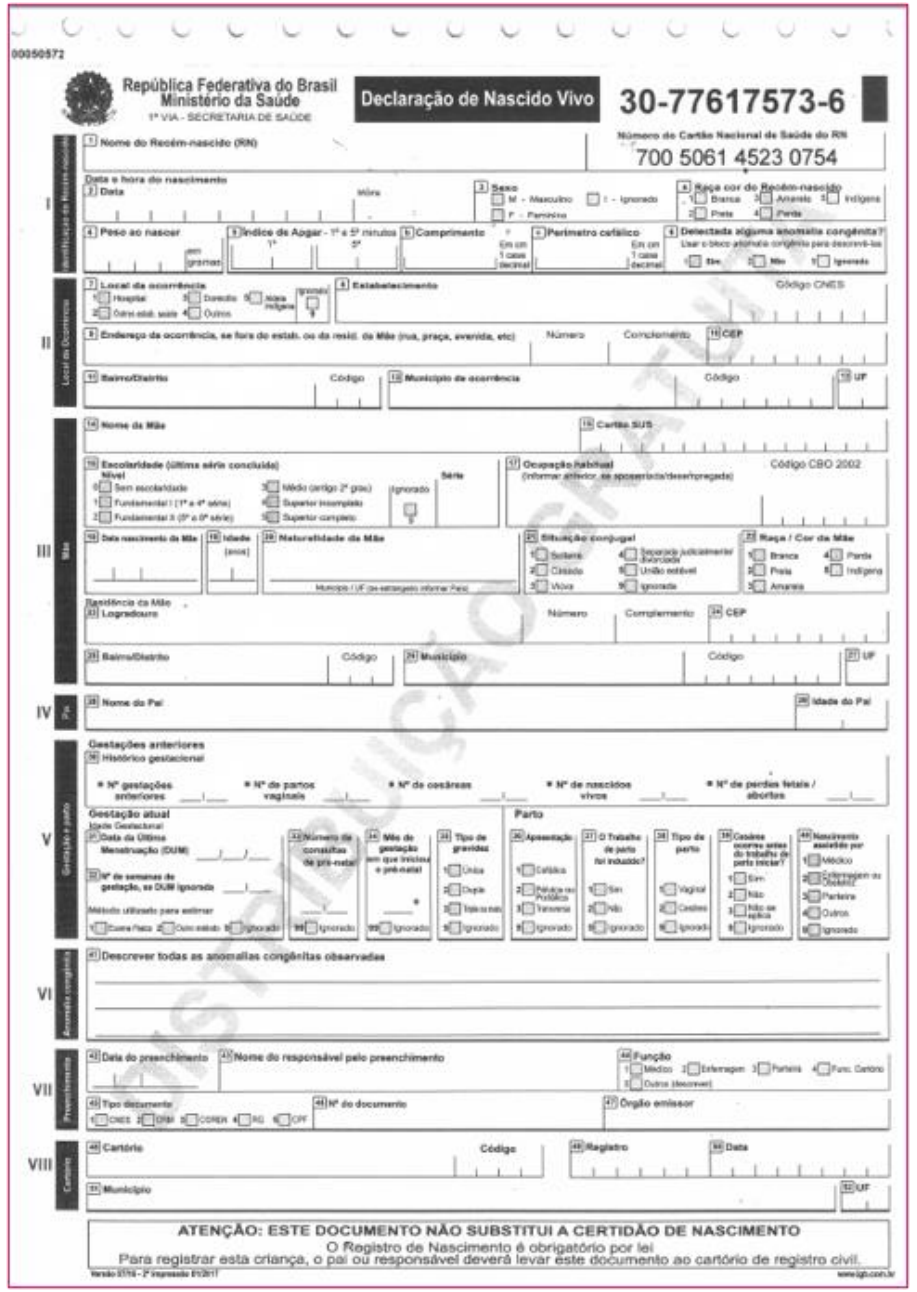
Tecnologia de visão computacional para vinculação de arquivos e **extração de informações de documentos oficiais**.

Aplicações das estatísticas:

- Desenvolvimento de **projeções populacionais**
- Elaboração de estimativas de **população em idade escolar**
- Planejamento de localização de **equipamentos de saúde**
- Análise de **tendências demográficas regionais**

ATUALIDADE | RELEVÂNCIA | CUSTO / BENEFÍCIO

ADEQUAÇÃO SISTEMAS TECNOLÓGICOS



Dados de telefonia móvel para análise de mobilidade

Metodologia permite análises precisas, desagregadas e atualizadas sobre deslocamentos populacionais para o planejamento de transportes e infraestrutura, especialmente em regiões sem pesquisas de origem e destino

MATRIZ DE VIAGENS

Por meio da identificação das **origens e destinos** de unidades de celular, é possível mapear **fluxos, duração, distância, número médio, concentrações de viagens** e dinâmicas dentro e entre munícipios.

AVALIAÇÃO DE INTERVENÇÕES

Comparação de matrizes de viagens antes e após intervenções, usando métodos causais para avaliar impacto de nova infraestrutura, equipamentos ou eventos na mobilidade urbana e regional.

Exemplos internacionais:

Statistics Netherlands (CBS): utiliza dados de telefonia móvel para analisar padrões de mobilidade, estimar distâncias médias percorridas, tempos de deslocamento e fluxos populacionais.

Instituto Nacional de Estadística (INE) da Espanha: utiliza esses dados para estatísticas de TURISMO; mede presença e deslocamentos de visitantes de forma contínua e com desagregação territorial.

ATUALIDADE | RELEVÂNCIA | ACESSIBILIDADE | PRECISÃO ???

IA Generativa para disseminação de conteúdo

Plataforma baseada em IA Generativa para produção automatizada de conteúdo sobre o Estado de São Paulo.
Utiliza bases de dados e análises produzidas pelo SEADE como fonte primária.

Reunião de bancos de dados e textos analíticos e metodológicos do SEADE

Processamento por IA para consulta, análise e interpretação dos dados por modelos de linguagem

Geração de conteúdo

Criação automatizada de relatórios, análises e visualizações

Disseminação personalizada às necessidades dos usuários

Tecnologia para transformar dados e textos em respostas a perguntas de usuários

Centro de Ciência de Dados para Estatísticas Públicas (CCDEP)

PROJETO FAPESP

- Início em outubro de 2024 e duração de 5 anos
- SEADE como instituição sede

PARCERIAS

- Instituições de pesquisa: FFLCH, POLI, IME e ICMC da USP, Unesp, FGV, FBSP
- Órgãos de governo: Sefaz, CPTM, STM, SSP e Conselho Estadual de Educação.

OBJETIVOS

- Utilizar dados de alta frequência, em grande volume e não estruturados para indicadores de políticas públicas
- Desenvolver novas ferramentas computacionais para coleta, processamento, análise e disseminação.
- Disseminar conhecimento na administração pública.

LINHAS DE PESQUISA:

1. **ATIVIDADE ECONÔMICA**: **Registros fiscais para indicadores econômicos (NFe)**, indicadores setoriais e regionais, sistema multiagente para coleta de textos e uso de IA Generativa

2. **MERCADO DE TRABALHO**: Coleta com URA para caracterizar ocupados, desocupados, inativos, **coleta de anúncios de emprego** e classificação textual de competências profissionais com ML

3. **MOBILIDADE URBANA**: Indicadores a partir de dados de telefonia móvel, matrizes origem-destino para regiões metropolitanas e integração com dados de bilhetagem eletrônica

4. **SEGURANÇA PÚBLICA**: Metodologias para **rotinas de PLN de registros criminais** e outros documentos e para **indicadores de população flutuante**, integrando com dados demográficos

5. **TECNOLOGIAS EM ENGENHARIA DE DADOS E DE SOFTWARE**: P&D de plataforma em Data Mesh, aprimorar modelos de coleta em plataformas web e classificação de conteúdos com novas ferramentas e uso de IA generativa

Obrigado!

Carlos Torres Freire
carlosfreire@seade.gov.br

SEADE
Fundação Sistema Estadual de Análise de Dados

GOVERNO DO ESTADO DE SÃO PAULO