Teoria e Métodos no Uso de Registros Administrativos na Pesquisa Aplicada em Políticas Públicas

Ricardo Ceneviva

Métodos Computacionais

Julho 2025

Objetivos da Apresentação

- Definir o que são registros administrativos (RA) e contextualizar seu papel em políticas públicas.
- Apresentar os métodos centrais para tratamento, integração e proteção de RA.
- Ilustrar aplicações com dois estudos de caso paradigmáticos (Chetty *et al.*, 2014; Barreto *et al.*, 2021).
- Discutir governança, ética e sustentabilidade dos dados na administração pública.

Roteiro da Aula

1. Ecossistema de RA
2. Metodologias de Integração & Proteção
3. Estudos de Caso
   - Chetty *et al.*, 2014
   - Barreto *et al.*, 2021
4. Governança, Ética & Sustentabilidade
5. Agenda Estratégica para o Futuro

RA: Definição & Tipologia

Definição

Dados coletados rotineiramente por governos para finalidades operacionais (cadastro, fiscalização, prestação de serviços).

Principais Tipos

- *Cadastrais*: identificação de pessoas, empresas, imóveis.
- *Transacionais*: eventos recorrentes (benefícios, internações).
- *Financeiros*: arrecadação, folha de pagamento.
- *Serviços*: registros de escola, saúde, segurança.

Exemplos Brasil: Cadastro Único, Censo Escolar, RAIS/CAGED, DATASUS.

Ciclo de Vida dos RA

1. **Coleta**: digitados na ponta (escolas, postos de saúde, repartições).
2. **Processamento**: validação e deduplicação em sistemas centrais.
3. **Armazenamento**: bases legadas ou *data lakes* na nuvem.
4. **Integração**: preparação e *linkage* para pesquisa ou avaliação.
5. **Arquivamento/Descarte**: períodos de retenção e destruição segura.

Vantagens dos Registros Administrativos

- Cobertura quase censitária $\rightarrow$ poder estatístico elevado.
- Longitudinalidade natural $\rightarrow$ acompanha indivíduos ao longo do tempo.
- Custo marginal de pesquisa quase nulo $\rightarrow$ aproveita dados já coletados.
- Precisão em variáveis sensíveis (ex.: renda tributária, óbitos).

Limitações e Desafios

- **Vieses de cobertura**: populações rurais e informais sub-representadas.
- **Qualidade heterogênea**: erros de digitação, códigos obsoletos.
- **Rigidez conceitual**: variáveis criadas para gestão, não para pesquisa.
- **Barreiras de acesso**: sigilo, fragmentação institucional.

Metodologia — Record Linkage

- **Determinístico**: chave única (CPF, NIS) $\rightarrow$ simples, mas dependente de completude.
- **Probabilístico**: Fellegi–Sunter (CIDACS-RL) $\rightarrow$ flexível, requer parâmetros e validação.
- Indicadores de qualidade: taxa de falsos-positivos/negativos, *precision*, *recall*.

PETs & Proteção de Privacidade

- **PPRL**: linkage com hashes/Bloom filters evitando troca de identificadores.
- **Privacidade Diferencial**: garante anonimato matemático (-DP) adicionando ruído.
- **Dados Sintéticos**: conjuntos artificiais que preservam estatísticas essenciais.
- **Aprendizado Federado**: modelos treinados sem mover microdados.

Qualidade & Integração de Dados

- Padronização de variáveis (códigos, datas, nomes).
- Imputação de valores faltantes (regressão, múltipla imputação).
- Metadados e *data profiling* para auditoria de consistência.
- Documentação reprodutível (R Markdown, Jupyter, data dictionaries).

Estudo de Caso 1 — Chetty et al. (2014)

Objetivo

Medir mobilidade intergeracional de renda nos EUA e identificar determinantes locais.

Dados — Registros fiscais (IRS) + dados de censo; 40 milhões de crianças (1980-1985).
Metodologia

- Linkage determinístico via Social Security Number.
- Regressão rank-rank (renda pais $\times$ renda filhos aos 35 anos).

Insights do Caso 1

- Mobilidade varia até $3\times$ entre condados $\rightarrow$ fenômeno fortemente local.
- Fatores correlacionados: segregação racial, qualidade escolar, capital social.
- Implicação: políticas territoriais (vouchers de moradia, zonings) podem aumentar oportunidades.

Estudo de Caso 2 — Barreto et al. (2021)

Coorte 100 Milhões de Brasileiros (CIDACS/Fiocruz)

Integra Cadastro Único (programas sociais) aos sistemas de saúde (SIM, SINASC, SINAN), cobrindo 131 milhões de indivíduos (2001-2018).

Método

- Record linkage probabilístico de alta escala (CIDACS-RL).
- Avaliação de impacto usando modelagem de painel.

Insights do Caso 2

- Programa Bolsa Família associado à redução de mortalidade materna e de tuberculose.
- Demonstra viabilidade de RA em países de renda média para estudos populacionais.
- Geração de infraestrutura de dados permanentes para futuras avaliações.

Lições Comparativas — Chetty vs. Barreto

- Escala nacional possível em contextos de alta (EUA) e média renda (Brasil).
- Identificadores fortes (SSN vs. NIS) e infraestrutura influenciam precisão do linkage.
- Necessário entender vieses de cobertura para interpretar evidências.

Governança de Dados — Five Safes

1. Safe Projects — projetos de interesse público legítimo.
2. Safe People — pesquisadores qualificados e credenciados.
3. Safe Data — dados anonimizados/tratados.
4. Safe Settings — ambientes controlados (laboratórios seguros, VDI).
5. Safe Outputs — checagem de resultados antes da divulgação.

Exemplo BR: Sala Segura do CIDACS atende às cinco dimensões.

Modelos de Stewardship & Data Spaces

- **Data Stewardship**: papéis claros de custodians e trustees para gerenciar acesso e uso.
- **Espaços de Dados da UE**: EHDS (saúde) $\rightarrow$ compartilhamento federado entre países.
- Estado como orquestrador: normas, infraestrutura e mecanismos de incentivo ao compartilhamento.

Sustentabilidade & Financiamento

- Custos fixos: servidores, segurança, equipe técnica.
- Modelos de co-financiamento público-privado (data collaboratives).
- Atenção a *data colonialism*: garantir benefícios locais e soberania.

Agenda Estratégica para o Futuro

- Padronização nacional de ID único (interoperabilidade CPF-NIS-CNS).
- Portais de consulta segura e aprendizado federado para análises inter-agência.
- Capacitação contínua de servidores e pesquisadores em ciência de dados.
- Incentivos regulatórios (Evidence Act BR) para compartilhamento responsável.

Encerramento & Discussão

- RA são pilar para políticas baseadas em evidências.
- Desafios: governança, qualidade, ética e financiamento.
- Perguntas, comentários e próximos passos.