IBGE

IBGE
Instituto Brasileiro de Geografia e Estatística

Escola Nacional de Ciências Estatísticas

Uso de pareamento estatístico para a criação de um cadastro de produtores rurais

Andrea Diniz da Silva
Programa de Pós-graduação em População, Território e Estatísticas Públicas

Seminário de Metodologia do IBGE
07 a 10 de novembro de 2016

Uso de pareamento estatístico para a criação de um cadastro de produtores rurais

Andrea, Diniz da Silva
National School of Statistical Science – ENCE/IBGE
Rua Andre Cavalcanti, 106
Rio de Janeiro, Brazil
andrea.diniz100@gmail.com

Flavio, Pinto Bolliger
The Food and Agriculture Organization of the United Nations – FAO
Rome, Italy
flavio.bolliger@fao.org

Jose Andre, de Moura Brito
National School of Statistical Science – ENCE/IBGE
Rua Andre Cavalcanti, 106
Rio de Janeiro, Brazil
jambrito@gmail.com

Resumo

1. Contexto e motivação

2. Construção do cadastro de produtores
    - 2.1 Fontes de dados
    - 2.2 Informações Disponíveis
    - 2.3 Estratégia de pareamento
    - 2.4 Métodos de pareamento

3. Discussão

1. Context

- A maior parte das estatísticas sobre a agropecuária é produzida pelo IBGE, portanto sob o Sistema Estatístico Nacional.
- Desde 1920 há censo agropecuário e desde 1938 o IBGE realiza pesquisas subjetivas.
- Censos decenais e pesquisas subjetivas formam o pilar das estatísticas oficiais sobre a gropecuárias.
- Para mudar o modelo de produção o IBGE está investindo na implantação de um sistema de pesquisas por amostragem probabilíestica (SNPA).
- Problema: não há, ainda, um cadastro de unidades (*Master Sampling Frame),* com boa qualidade.

1. Contexto

- Há vários cadastros disponíveis, permitindo o uso da técnica *Multiple Frame.*
    - Area: IBGE (boa qualidade).
    - List: IBGE, State Tax Files, Ministry of Agrarian Development (com limitações).
- Várias técnicas de pareamento (Record Linkage): nenhuma superior a despeito dos dados.
- Compar as técnicas, com dados reais é necessária.

2. Construção do Cadastro de Produtores

- Uma extensão do trabalho realizado em 2012.
    - Mais de 50 mil pares foram obtidos, comparando Censo Agro 2006, CNEFE 2010 e Cadastro de Empresas (CEMPRE).
- Inovações do trabalho:
    - Cinco fontes de dados: Censo Agro 2006, CNEFE 2010, CEMPRE 2015, Secretarias de Estado de Fazenda 2016 e MDA 2016.
    - Outros métodos: árvore de decisão e SVM.

2.1 Fontes de Dados

Table 1: *Number of records available for linkage by source*

State	Source				
	IBGE			State	MDA
	Census of Agriculture (2006)	Addresses File (2010)	Business Register (2014)	Tax Files (2016)	Family Farming (2016)
Ceara	381,017	84,954	979	509	606,383
Mato Gosso do Sul	64,864	68,453	1,262	123,104	29,554
Paraíba	167,286	88,092	298	649	215,008
Maranhão	287,039	35,761	647	13,475	410,882
Santa Catarina	170,913	150,043	1,803	92,338	122,527

- Censo agro tem melhor cobertura que o demográfico.
- CEMPRE tem poucos produtores.
- Sec Fazenda pode incluir endereço do escritório.
- MDA pode ter mais de um produtor da mesma unidade.

2.2 Informações disponíveis

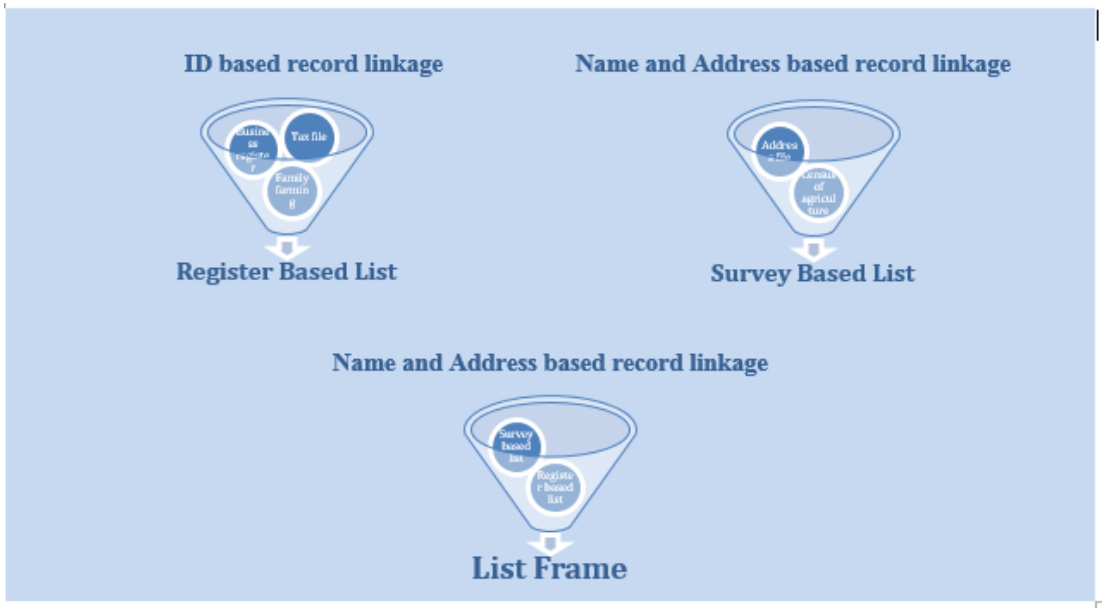
Table 2: *Main information available by source*

Variable	Source								
	State Tax File					IBGE			MDA
	Ceara	Mato Gosso do Sul	Paraiba	Maranhão	Santa Catarina	Census of agriculture	Address file	Business Register	Family Farming
Name	x	x	x	x	x	x	x	x	x
ID on Tax File*	x	x	x	x	x	-	-	x	x
Activity Code	x	-	x	-	x	x	-	x	-
Activity Description	x	-	x	-	x	x	-	x	-
Address (street, number etc.)	x	x	x	x	x	x	x	x	-
Owners: name	x	-	x	x	-	x	-	-	-
Owners: ID on Tax File	x	-	x	-	-	-	-	-	-
Owners: address	x	-	-	-	-	-	-	-	-

* CPF ou CNPJ

- Basicamente nome e endereço.
- Há muito mais informações disponíveis, porém a política e os protocolos para acesso dificultam a obtenção.

2.3 Estratégia de Pareamento



- Três procedimentos inependentes
  - Registro <-> Registro.
  - Pesquisa<-> pesquisa.
  - Pesquisa <-> registro.

2.4 Métodos e ferramentas de pareamento

- Duas classes de modelos teóricos
    - Estocástico: FS modelo de decisão.
    - Não-estocásticoc: Árvore de Decisão e SVM.
- Blocagem: Município ou Estado?
- Função de comparação: Jaro-Winkler.
- Linguagem de programação R.

3. Discussão

- Uso de dados de fontes com produção regular aumenta as chances de manutenção.
- Ainda não há uma prática de compartilhar dados, mesmo entre órgão públicos.
- Há um grande número de técnicas de pareamento, porém a escolha de uma delas depende das condições de pareamento: estrutura dos dados e recursos disponíveis.
- O estudo se limita a cinco Unidades da Federação, porém permite extensão para as demais.

Obrigada