Working with Administrative Data

The CPR-Government of Andhra Pradesh (GoAP) Partnership

CENTRE FOR POLICY RESEARCH

Neelanjan Sircar

Use-Cases

Policy Problems and Academic Value Propositions

- ***Surveys and Scheme Evaluation***
  - Populations Database with Key Characteristics
  - Avoids Challenges of Sampling Frame Creation (NSS Debate)
- ***Monitoring Revenue (Example: Excise Tax)***
  - Choosing the Correct Statistical Models
  - Sales and Deviations are Good Dependent Variables
- ***Fraud Detection (Example: Commercial Tax)***
  - Transactions, Registration, and Scrutiny
  - Real-Time Economic Modeling (Hard Problem)

CENTRE FOR
POLICY
RESEARCH

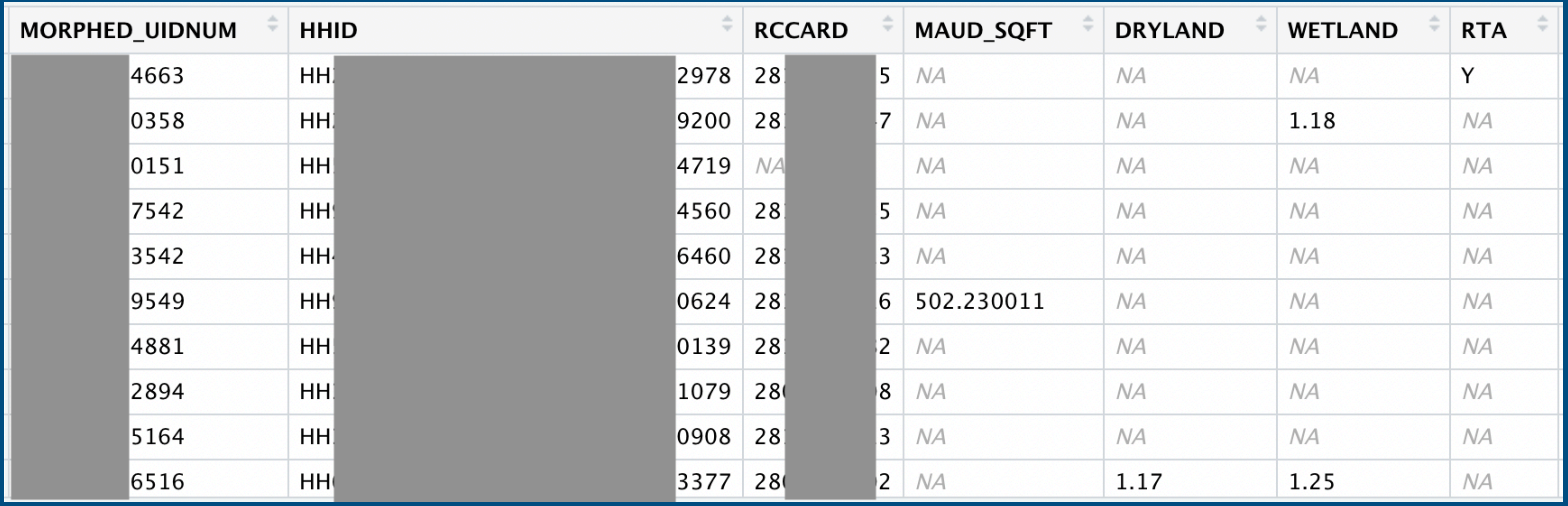
Surveys and Scheme Evaluation

A Look at the AP Urban Youth Survey

Administrative Population Data

The GSWS Data Structure

- 1.65cr households, 90%+ population coverage
- Databases used for beneficiaries (parameters for inclusion)
- Researcher never should observe UID/Aadhar — Masking Protocols

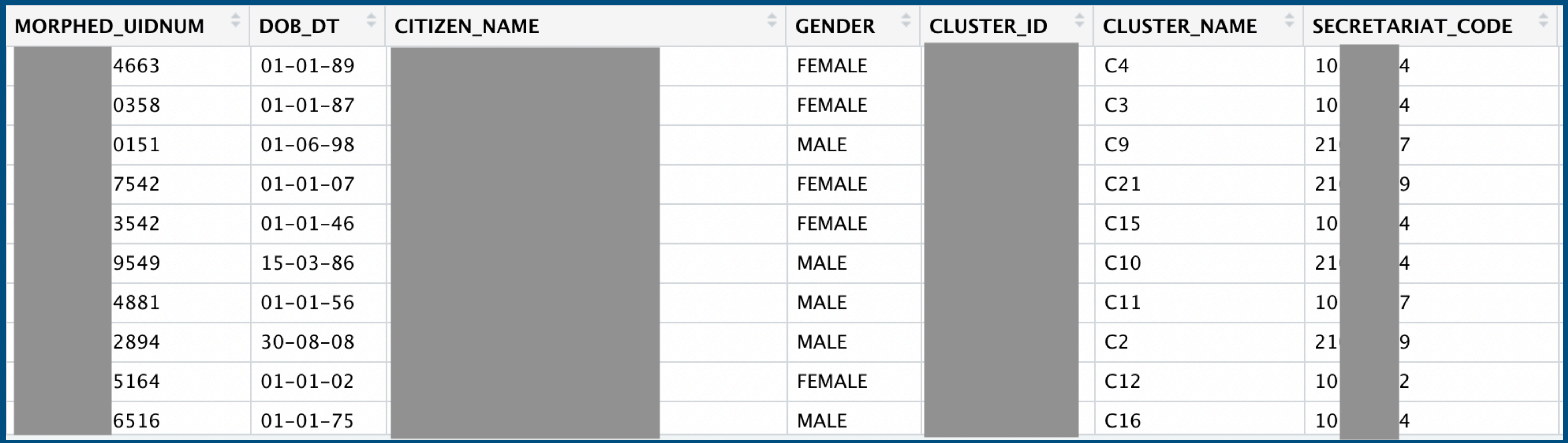


MORPHED_UIDNUM	HHID	RCCARD	MAUD_SQFT	DRYLAND	WETLAND	RTA
[redacted]4663	HH[redacted]2978	28[redacted]5	NA	NA	NA	Y
[redacted]0358	HH[redacted]9200	28[redacted]7	NA	NA	1.18	NA
[redacted]0151	HH[redacted]4719	NA[redacted]	NA	NA	NA	NA
[redacted]7542	HH[redacted]4560	28[redacted]5	NA	NA	NA	NA
[redacted]3542	HH[redacted]6460	28[redacted]3	NA	NA	NA	NA
[redacted]9549	HH[redacted]0624	28[redacted]6	502.230011	NA	NA	NA
[redacted]4881	HH[redacted]0139	28[redacted]2	NA	NA	NA	NA
[redacted]2894	HH[redacted]1079	28[redacted]8	NA	NA	NA	NA
[redacted]5164	HH[redacted]0908	28[redacted]3	NA	NA	NA	NA
[redacted]6516	HH[redacted]3377	28[redacted]2	NA	1.17	1.25	NA

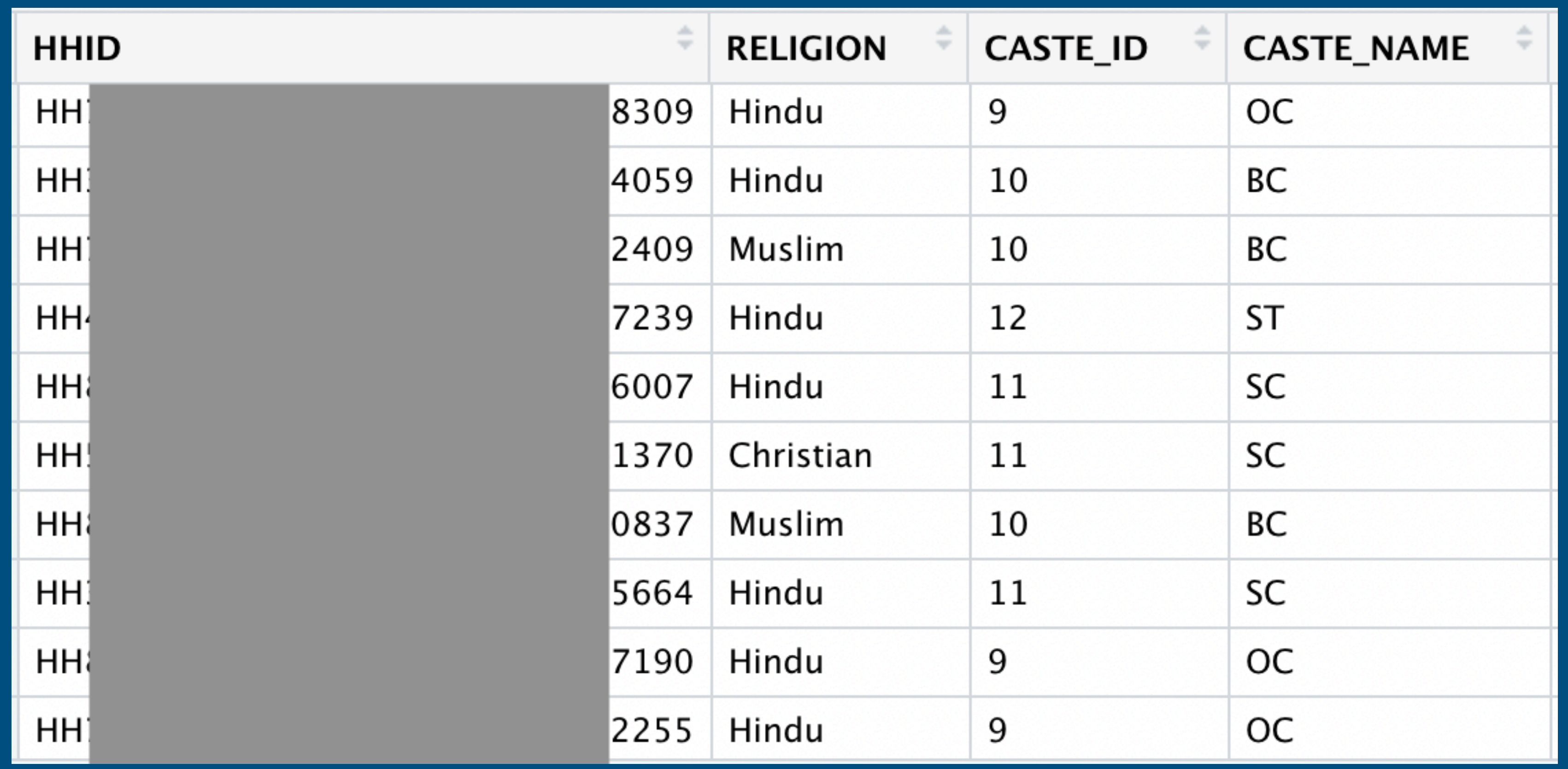
Identification in Databases

Attributes and Spatial Identifiers

- Beneficiary Schemes require individual/HH attributes (gender, age, identity)
- Approximately 15,000 secretariats, 3,300 individuals per secretariat
- Reach individuals through clusters, volunteers (no addresses)



MORPHED_UIDNUM	DOB_DT	CITIZEN_NAME	GENDER	CLUSTER_ID	CLUSTER_NAME	SECRETARIAT_CODE
[redacted]4663	01-01-89	[redacted]	FEMALE	[redacted]	C4	10[redacted]4
[redacted]0358	01-01-87	[redacted]	FEMALE	[redacted]	C3	10[redacted]4
[redacted]0151	01-06-98	[redacted]	MALE	[redacted]	C9	21[redacted]7
[redacted]7542	01-01-07	[redacted]	FEMALE	[redacted]	C21	21[redacted]9
[redacted]3542	01-01-46	[redacted]	FEMALE	[redacted]	C15	10[redacted]4
[redacted]9549	15-03-86	[redacted]	MALE	[redacted]	C10	21[redacted]4
[redacted]4881	01-01-56	[redacted]	MALE	[redacted]	C11	10[redacted]7
[redacted]2894	30-08-08	[redacted]	MALE	[redacted]	C2	21[redacted]9
[redacted]5164	01-01-02	[redacted]	FEMALE	[redacted]	C12	10[redacted]2
[redacted]6516	01-01-75	[redacted]	MALE	[redacted]	C16	10[redacted]4



HHID	RELIGION	CASTE_ID	CASTE_NAME
HH[redacted]8309	Hindu	9	OC
HH[redacted]4059	Hindu	10	BC
HH[redacted]2409	Muslim	10	BC
HH[redacted]7239	Hindu	12	ST
HH[redacted]6007	Hindu	11	SC
HH[redacted]1370	Christian	11	SC
HH[redacted]0837	Muslim	10	BC
HH[redacted]5664	Hindu	11	SC
HH[redacted]7190	Hindu	9	OC
HH[redacted]2255	Hindu	9	OC

AP Urban Youth Survey

The Power of Administrative Population Data

- Study Population: Men and Women ages 19-29 (hard to reach)
- **Focus:** Skilling, Mobility, Educational Outcomes, Scheme Penetration, Preferences
- Cross-referenced with GSWS data — population prediction, estimating biases

CENTRE FOR POLICY RESEARCH

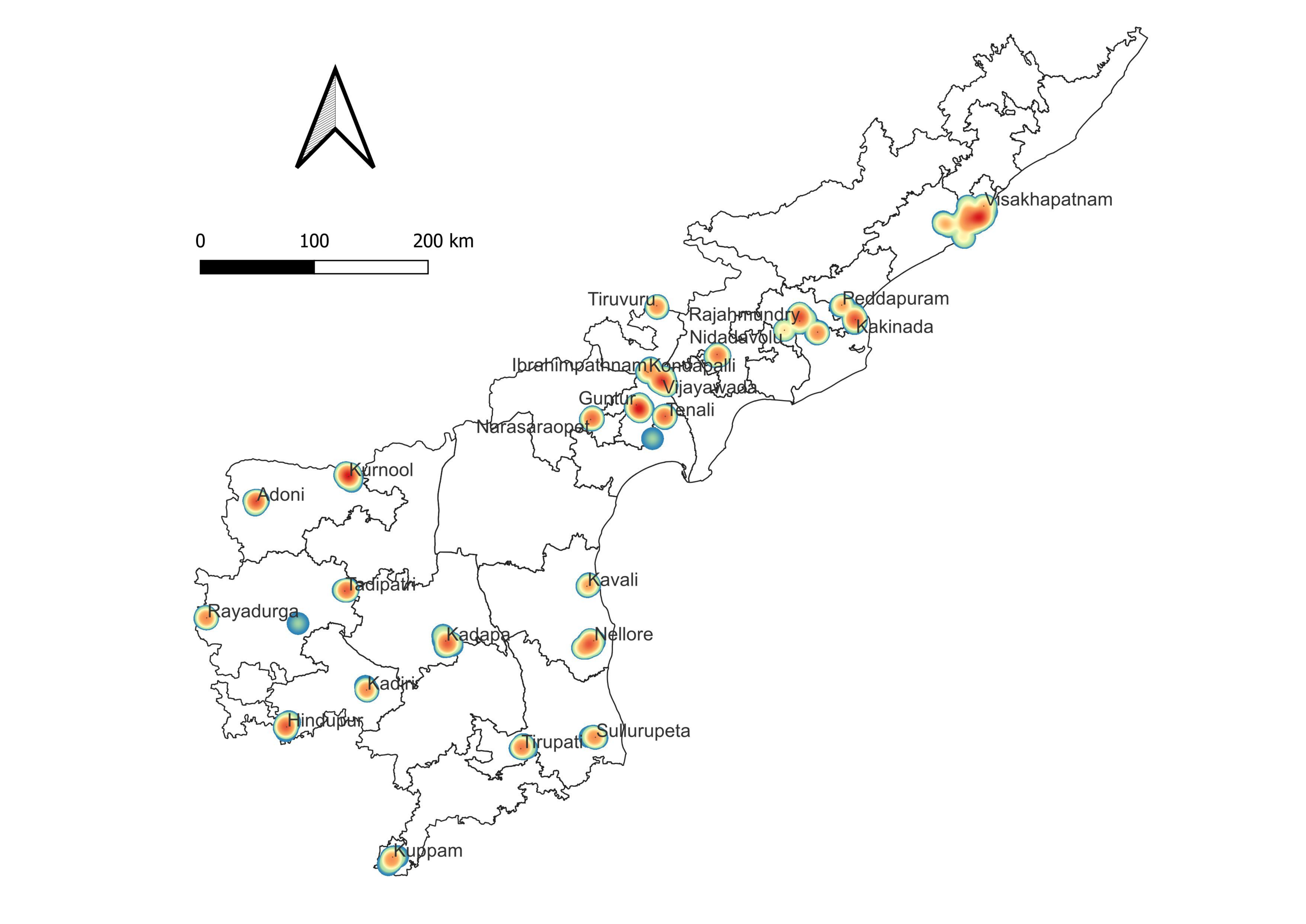
Andhra Pradesh
Urban Youth Survey

INDUS ACTION

Survey Overview

Spatially Representative

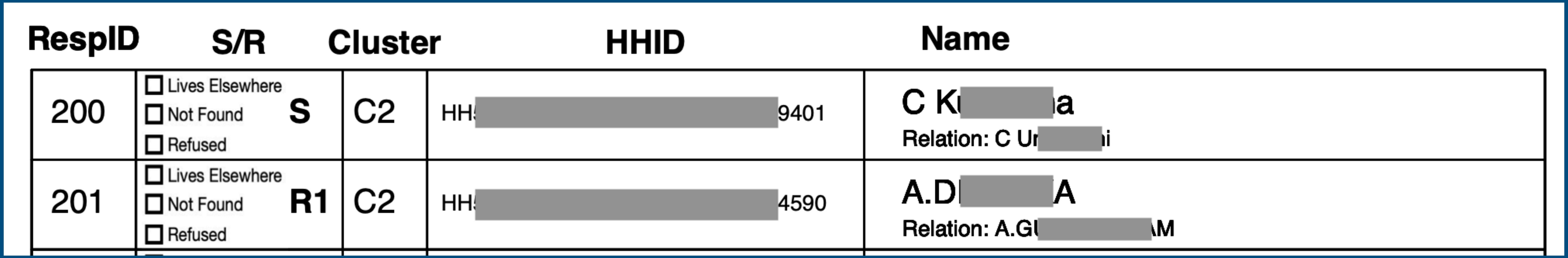
- Largest ever representative youth survey in India?
- GSWS linkage and sampling strategy
- 25 cities across 3 size classes: 4L+, 1-4L, <1L
- 4347 surveys over 330 secretariats



Robust Samples

Protocols, Findability, and Replacement

- Non-identifiable to researcher and a system of “double consent”
- No issues of listing or creating frames, reachable through local actors
- Higher response rate and replacement biases can be assessed



RespID	S/R		Cluster	HHID	Name
200	□ Lives Elsewhere □ Not Found □ Refused	S	C2	HH[redacted]9401	C K[redacted]a Relation: C U[redacted]i
201	□ Lives Elsewhere □ Not Found □ Refused	R1	C2	HH[redacted]4590	A.D[redacted]A Relation: A.G[redacted]M

CENTRE FOR POLICY RESEARCH

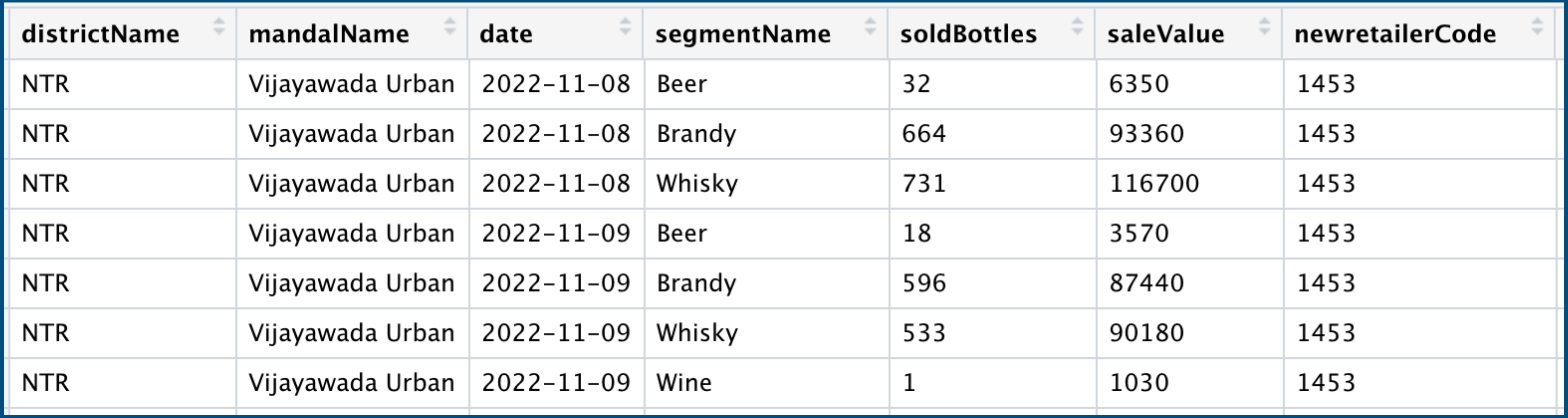
Revenue Monitoring

A Look at Work on Excise Tax

Working with Administrative Data

Alcohol Sales in Government Retail Outlets

- Data is for bookkeeping, not generated for any analytic purpose
- Assess feasibility and understand data generation
- The Role of Statistical Modeling

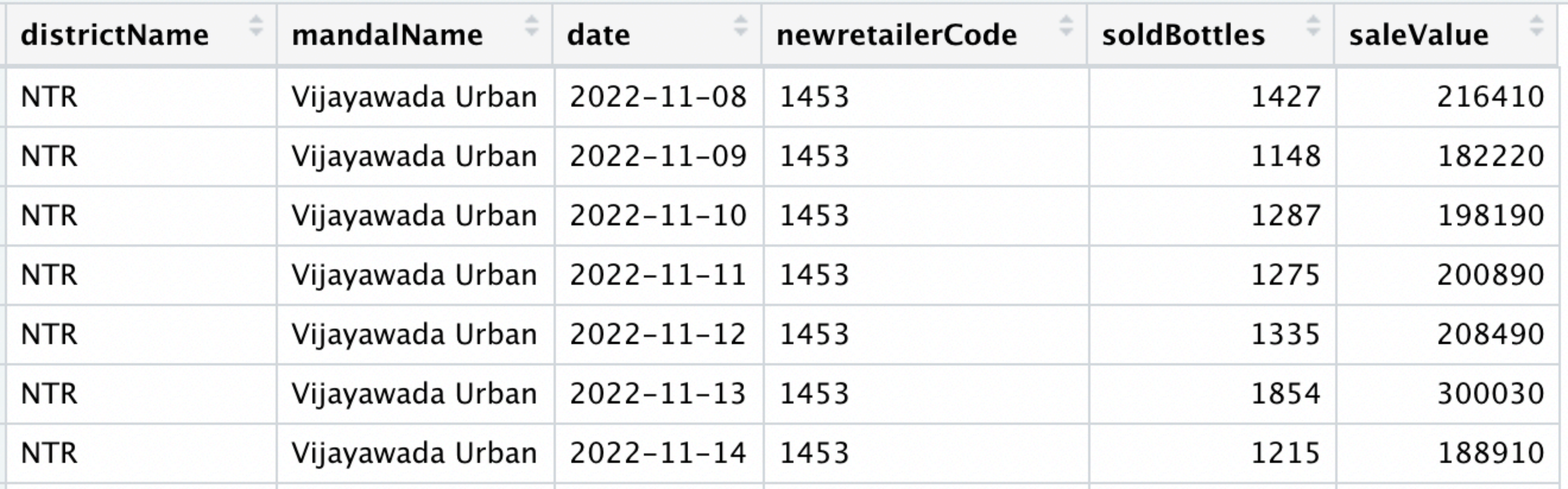


districtName	mandalName	date	segmentName	soldBottles	saleValue	newretailerCode
NTR	Vijayawada Urban	2022-11-08	Beer	32	6350	1453
NTR	Vijayawada Urban	2022-11-08	Brandy	664	93360	1453
NTR	Vijayawada Urban	2022-11-08	Whisky	731	116700	1453
NTR	Vijayawada Urban	2022-11-09	Beer	18	3570	1453
NTR	Vijayawada Urban	2022-11-09	Brandy	596	87440	1453
NTR	Vijayawada Urban	2022-11-09	Whisky	533	90180	1453
NTR	Vijayawada Urban	2022-11-09	Wine	1	1030	1453

A Modeling Approach

Tracking Revenues

- **Key Problem:** Characterize the revenue performance of various outlets
- Identify Revenue Benchmarks from Data!
- Define outcome (sale value) and pre-process data



districtName	mandalName	date	newretailerCode	soldBottles	saleValue
NTR	Vijayawada Urban	2022-11-08	1453	1427	216410
NTR	Vijayawada Urban	2022-11-09	1453	1148	182220
NTR	Vijayawada Urban	2022-11-10	1453	1287	198190
NTR	Vijayawada Urban	2022-11-11	1453	1275	200890
NTR	Vijayawada Urban	2022-11-12	1453	1335	208490
NTR	Vijayawada Urban	2022-11-13	1453	1854	300030
NTR	Vijayawada Urban	2022-11-14	1453	1215	188910

Building a Model

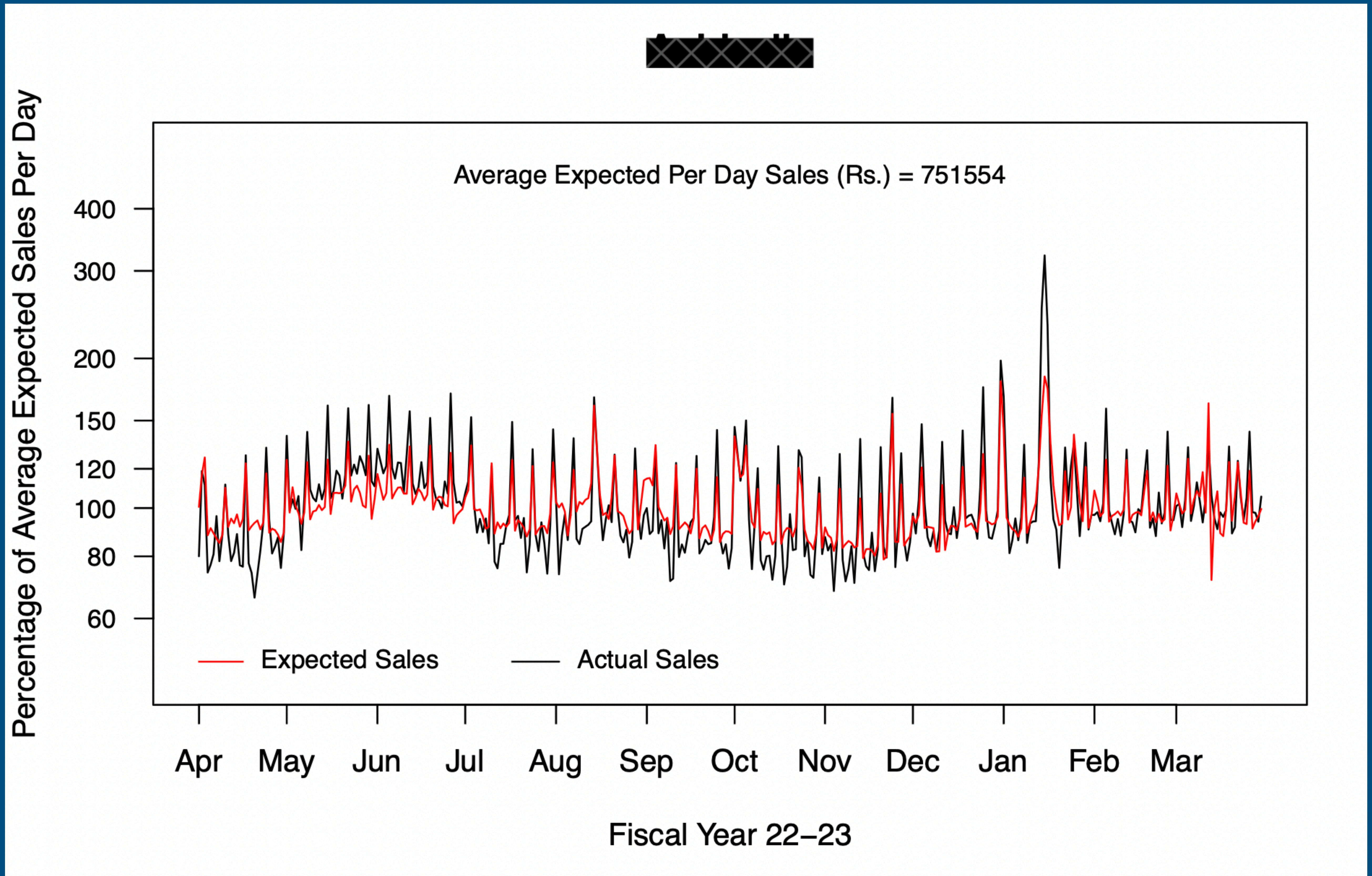
The Case for Statistics

CENTRE FOR POLICY RESEARCH

- Benchmark (red curve):
  $\log(y_{it}) = \mu_i + \alpha_t$
  where y_{it} = sales for retailer i on day t

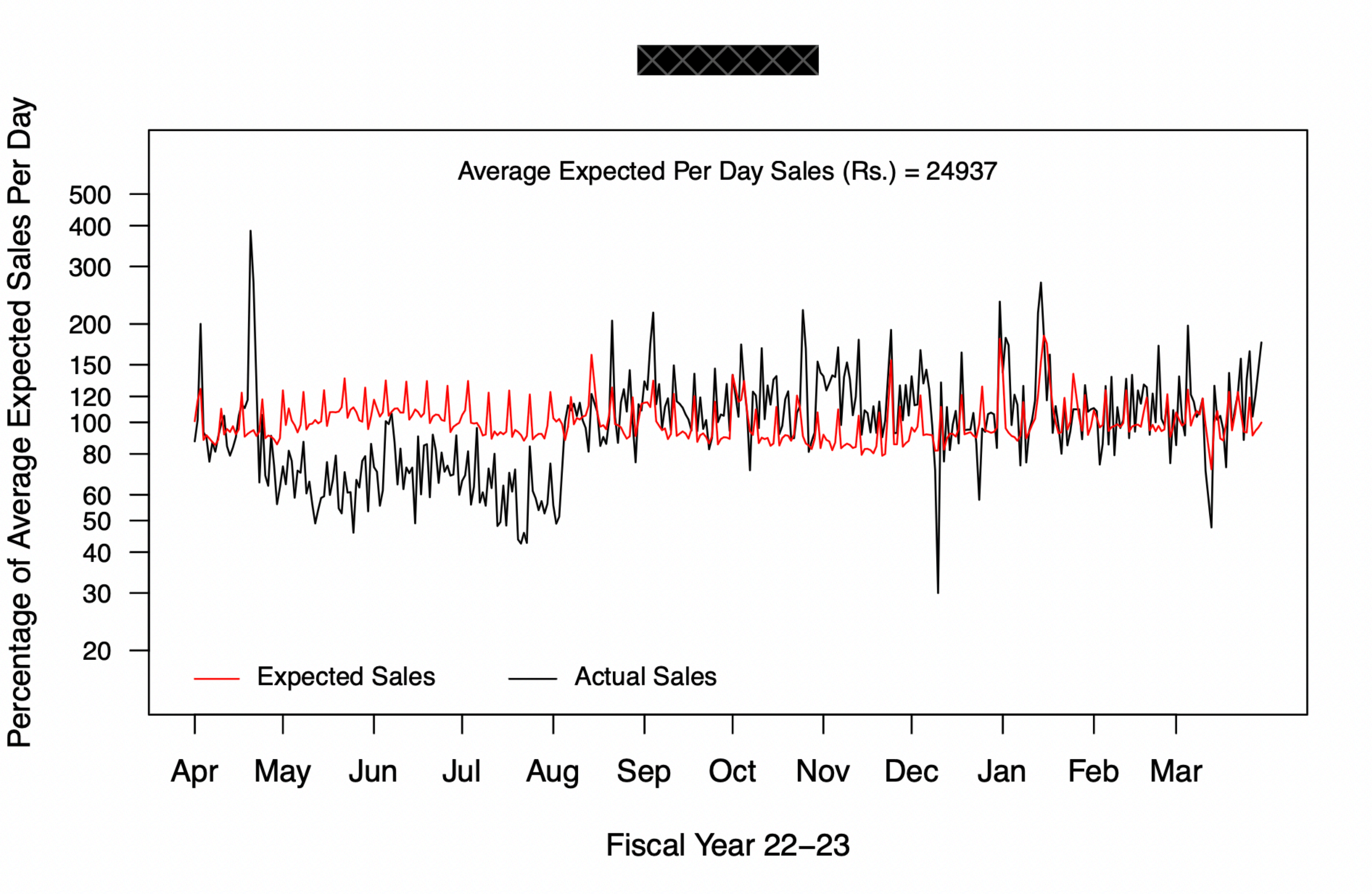
- Compare to pre-processed actual sales data (black curve)

- Fixed effects type model provide robustness in data sparse environment (machine learning?)

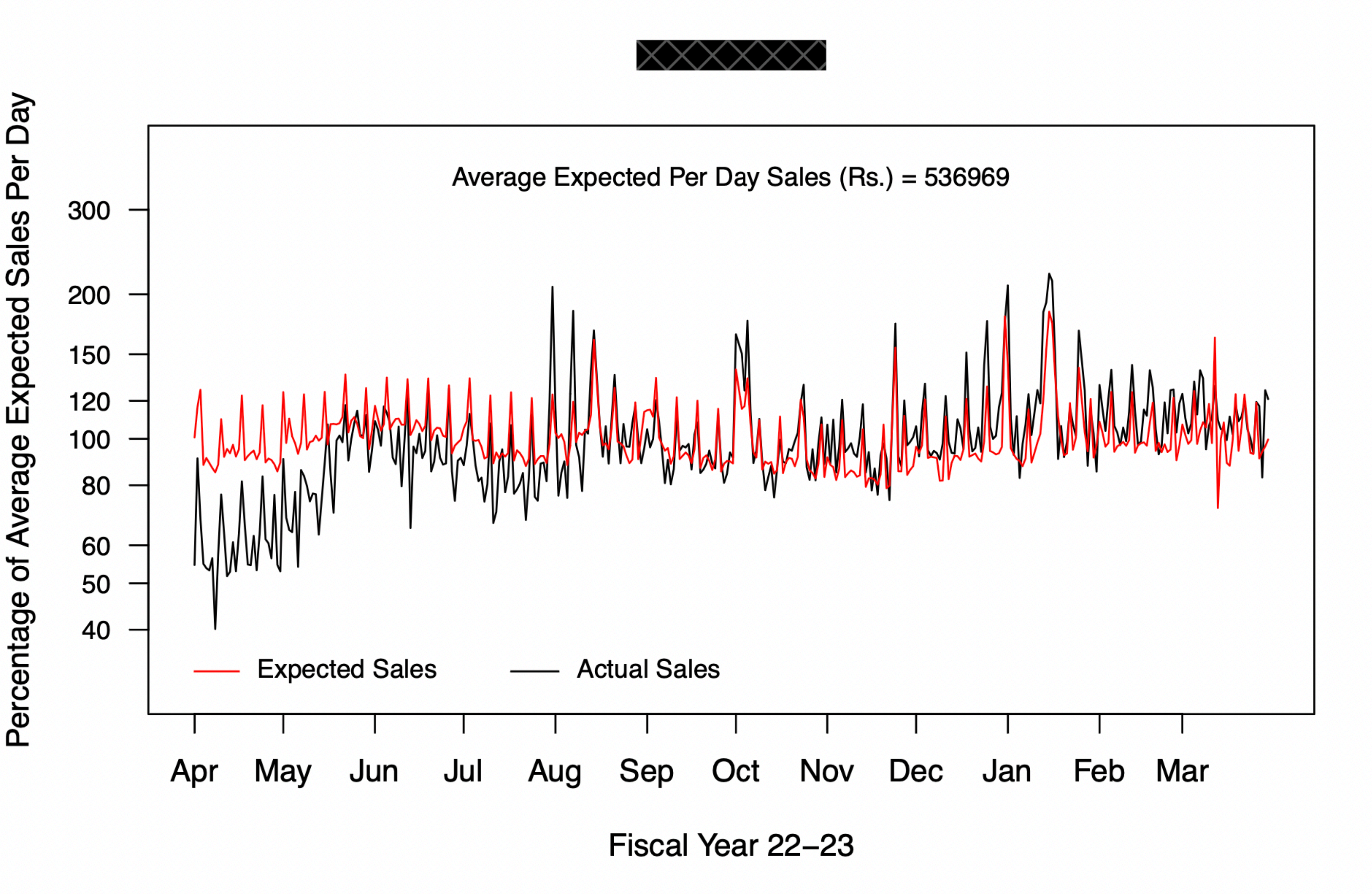


Communicating Models

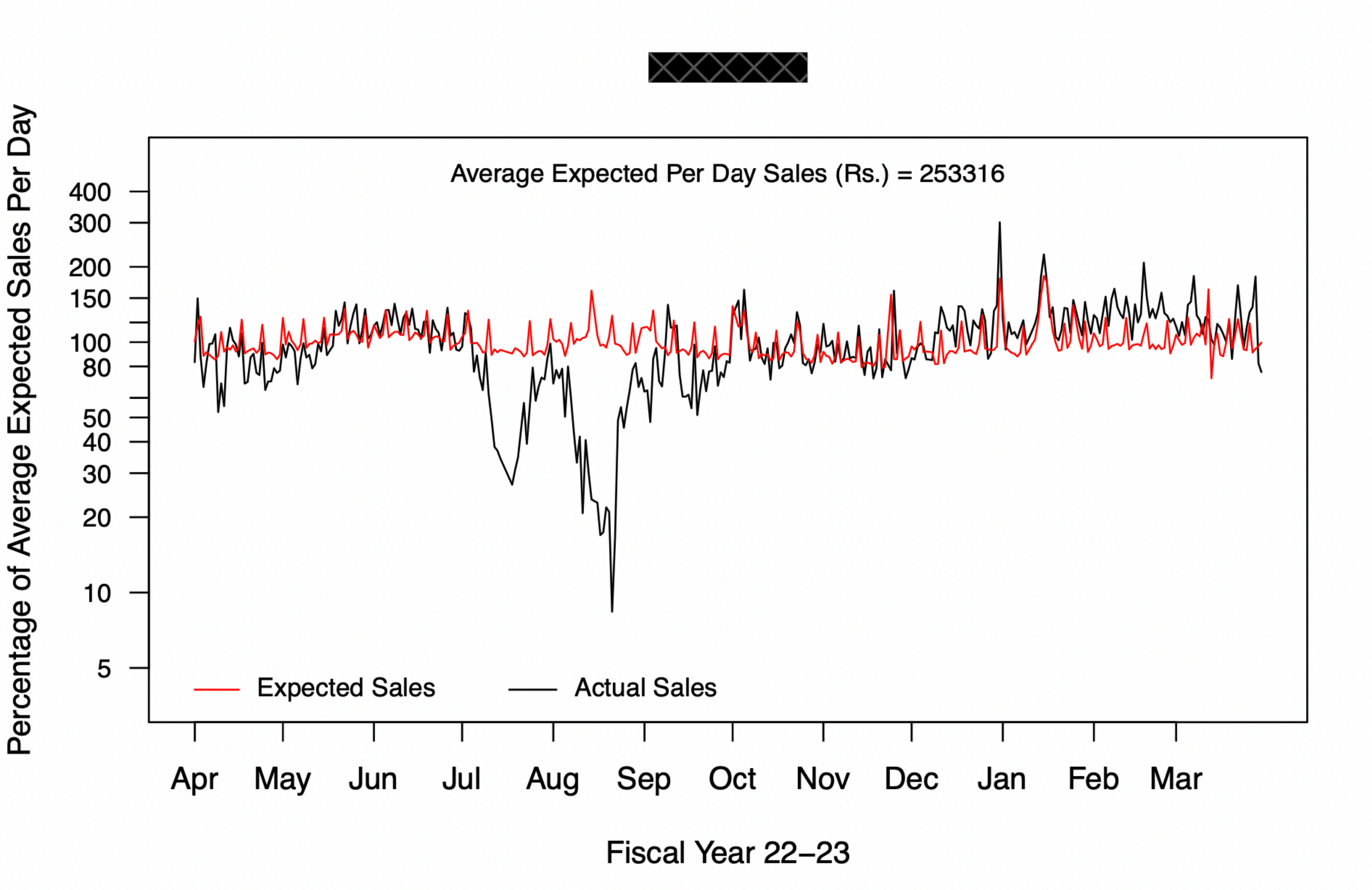
Visualizing Revenue Deviations



[redacted]



[redacted]



[redacted]

Using the Data

Developing Academic Questions

- ***Geo-Located, High-Frequency Data***
  - Event Analysis (weather, COVID, scheme delivery)
  - Cross-Reference with other data (GSWS)
- ***Policy-Relevant Questions***
  - Does it matter if the beneficiary is male or female?
  - A good measure of consumption behavior?
- ***Scrutiny***
  - More granular data to predict deviations
  - Incorporating field observations

CENTRE FOR POLICY RESEARCH

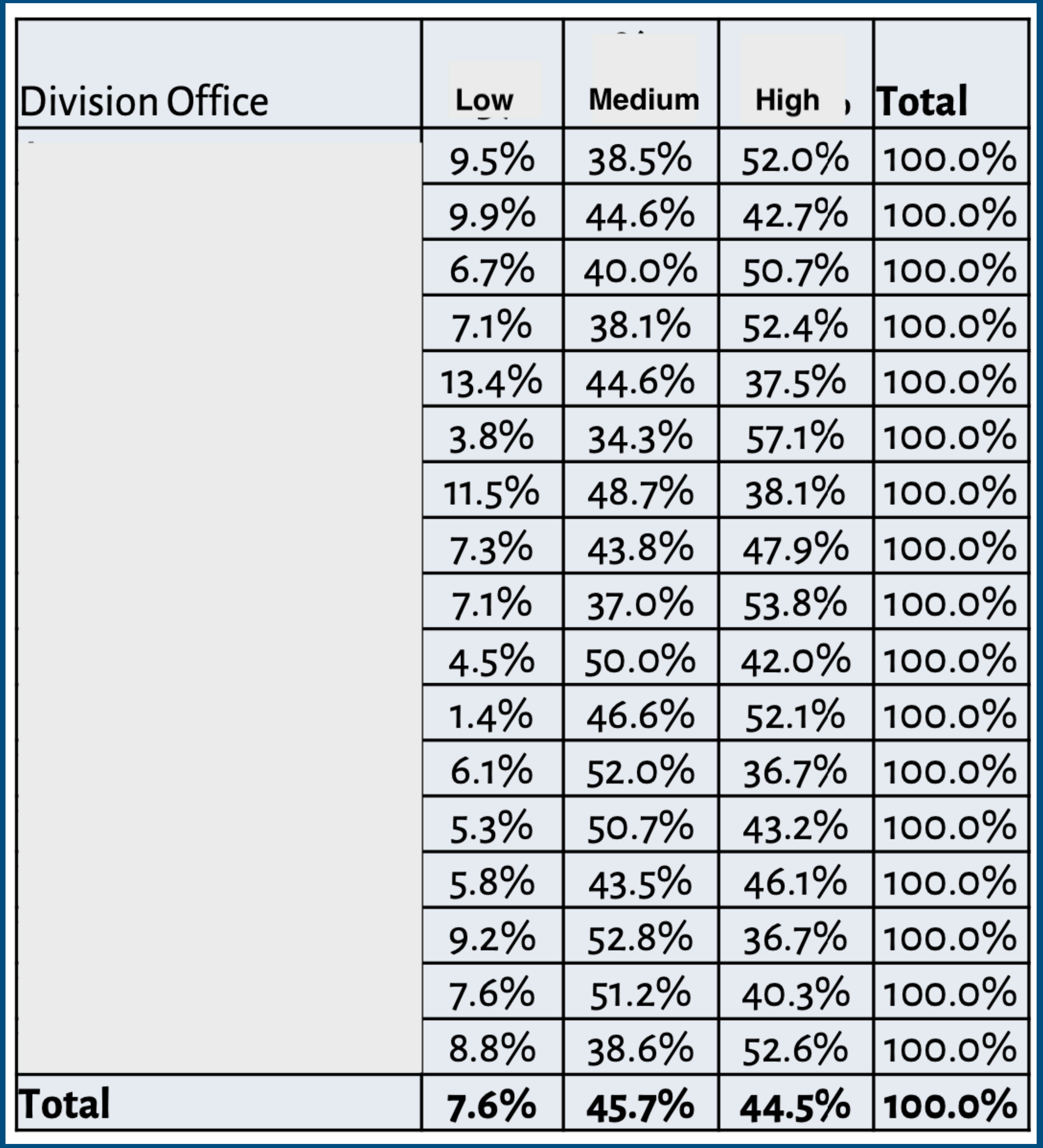
Fraud Detection

A Look at Work on Commercial Tax

Detecting Fraud

Overclaiming Tax Credit

- Receive tax credit (ITC) for inputs
- Must pay taxes on what you sell
- Offset assessed tax by paying with tax credits
- Major concern of fraud. One key analysis: unusually high ratio of assessed tax paid in tax credit.



Division Office	Low	Medium	High	Total
	9.5%	38.5%	52.0%	100.0%
	9.9%	44.6%	42.7%	100.0%
	6.7%	40.0%	50.7%	100.0%
	7.1%	38.1%	52.4%	100.0%
	13.4%	44.6%	37.5%	100.0%
	3.8%	34.3%	57.1%	100.0%
	11.5%	48.7%	38.1%	100.0%
	7.3%	43.8%	47.9%	100.0%
	7.1%	37.0%	53.8%	100.0%
	4.5%	50.0%	42.0%	100.0%
	1.4%	46.6%	52.1%	100.0%
	6.1%	52.0%	36.7%	100.0%
	5.3%	50.7%	43.2%	100.0%
	5.8%	43.5%	46.1%	100.0%
	9.2%	52.8%	36.7%	100.0%
	7.6%	51.2%	40.3%	100.0%
	8.8%	38.6%	52.6%	100.0%
Total	**7.6%**	**45.7%**	**44.5%**	**100.0%**

Ratio Credit/Total Tax by Division

Dealer Attributes

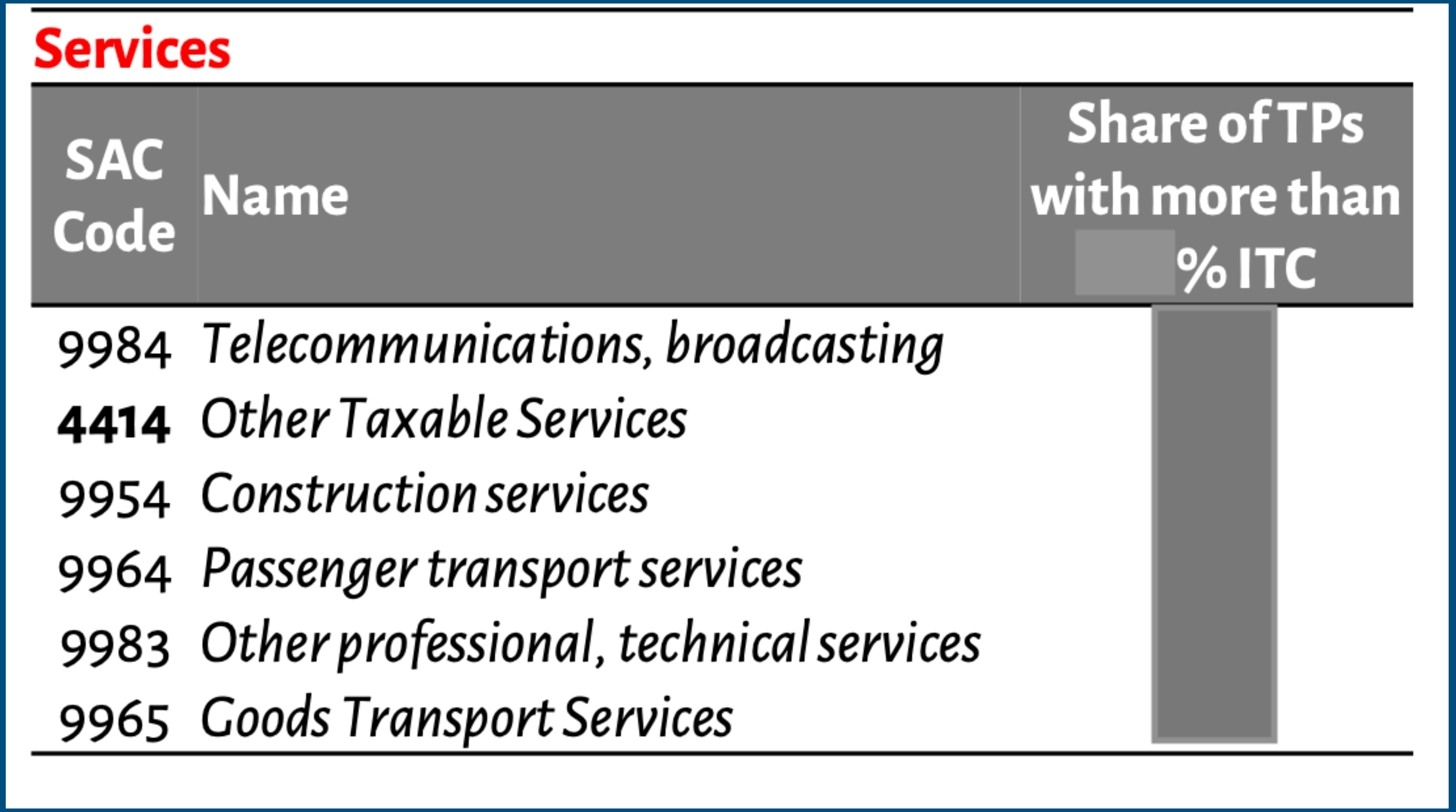
Predicting Overclaimed Credit

- Why would services have a high tax credit/total tax ratio?

- Look at attributes of company (from registration) and look for inconsistencies.

- Model the credit/total tax ratio and look for outliers

- Other attributes also important (e.g., age of firm, age of input firms)



Services

SAC Code	Name	Share of TPs with more than [redacted] % ITC
9984	*Telecommunications, broadcasting*	[redacted]
4414	*Other Taxable Services*	[redacted]
9954	*Construction services*	[redacted]
9964	*Passenger transport services*	[redacted]
9983	*Other professional, technical services*	[redacted]
9965	*Goods Transport Services*	[redacted]

Concluding Thoughts

Academic-State Government Partnership

- ***Still Largely a Consultant Space***
  - Make the case for quality and skill
  - Don’t enter without data (privacy and publishing) agreements
- ***Varied Questions***
  - High quality government data can be leveraged for better research
  - Lots of room for methodological innovation
- ***Room to Re-Imagine Development Planning and Economic Monitoring***
  - Requires expert understanding data structure and data generating process
  - Needs to remain apolitical