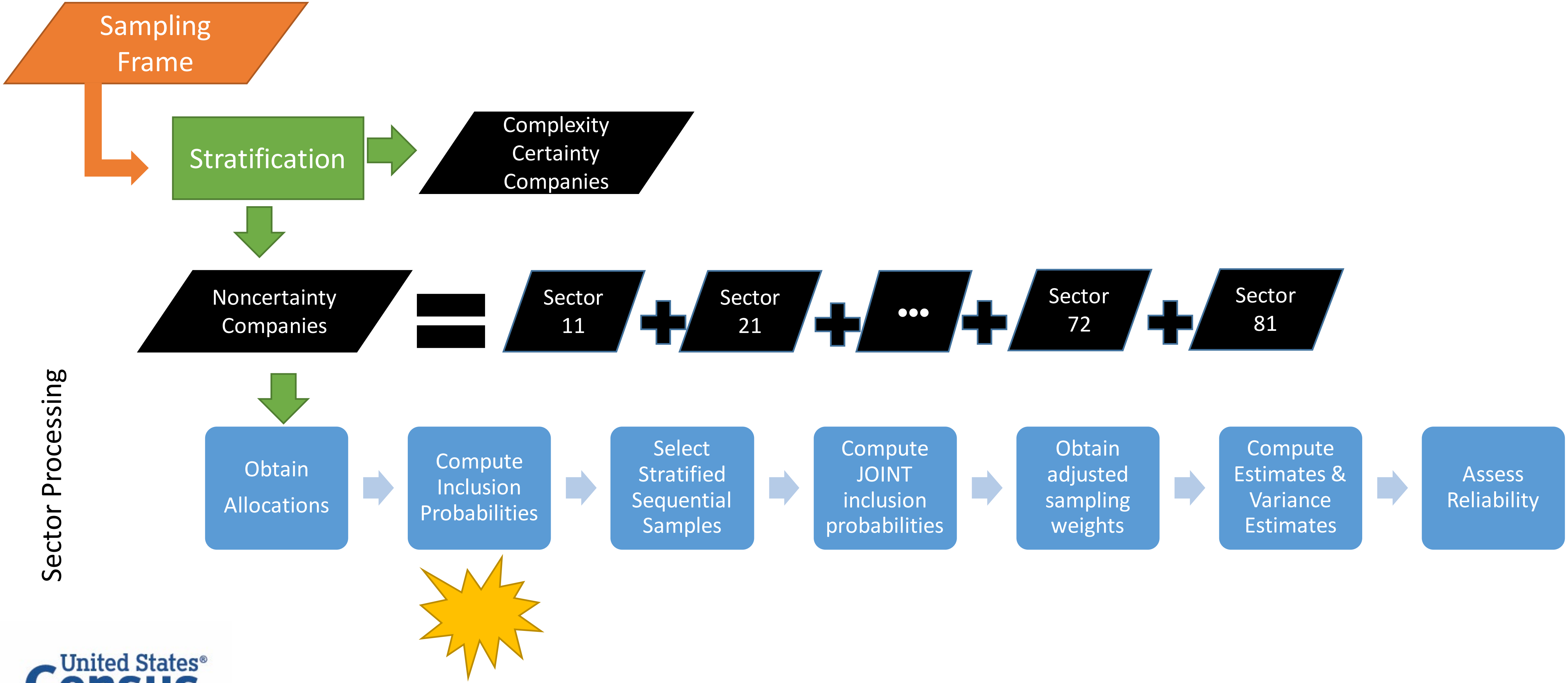
Winsorizing Low Inclusion Probabilities from an Unequal Probability Sample in a Multi-Purpose Annual Survey

Nikki Czaplicki and Yeng Xiong

Economic Statistical Methods Division, U.S. Census Bureau

Overview of AIES Sample Design Process

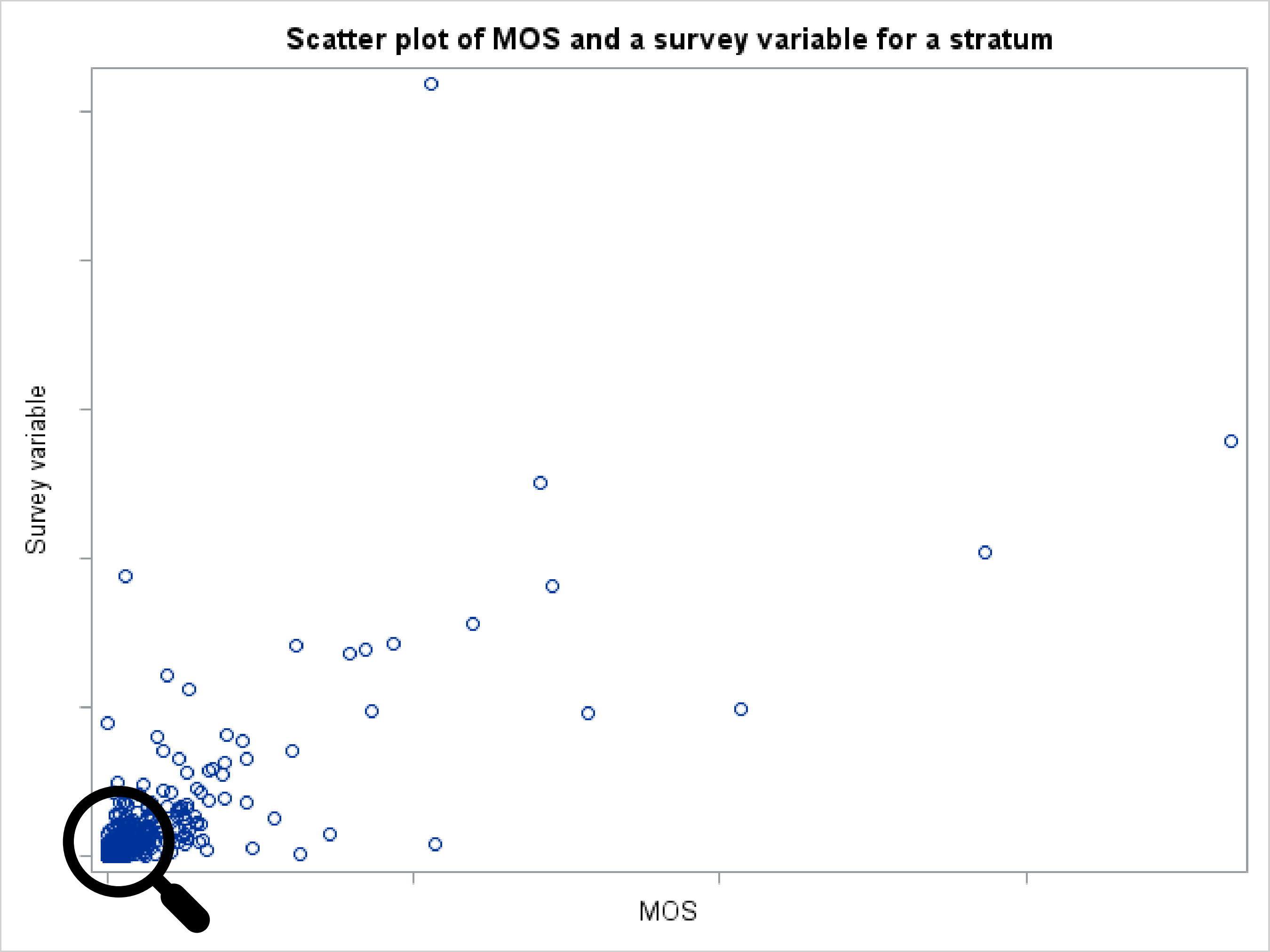


AIES sample inclusion probabilities

- Recall that the key estimates for AIES will be industry totals at both the national and subnational level.
- Strata are defined by the cross classification of three-digit NAICS by state (or balance of region).
- Like most economic data, the AIES data come from a skewed distribution with most of the total coming from a relatively small number of companies.
- Inclusion probabilities are assigned in a probability proportional to estimated size (PPES) design.

Probability Proportional to Size

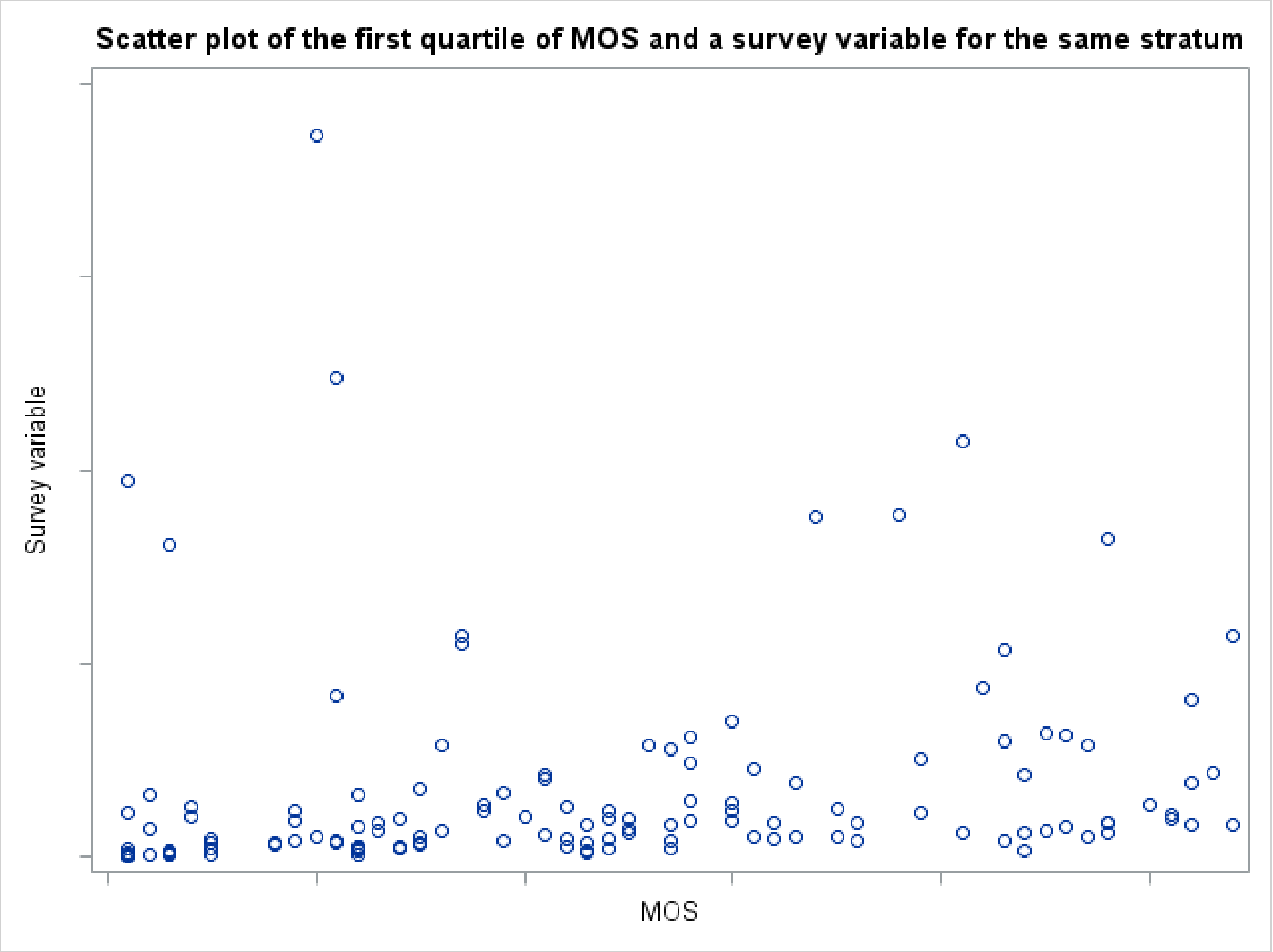
In PPS, the measure of size (MOS) is assumed to be linearly related to the survey variable.



Probability Proportional to Size

At the lower end of the distribution, the linear relationship no longer holds.

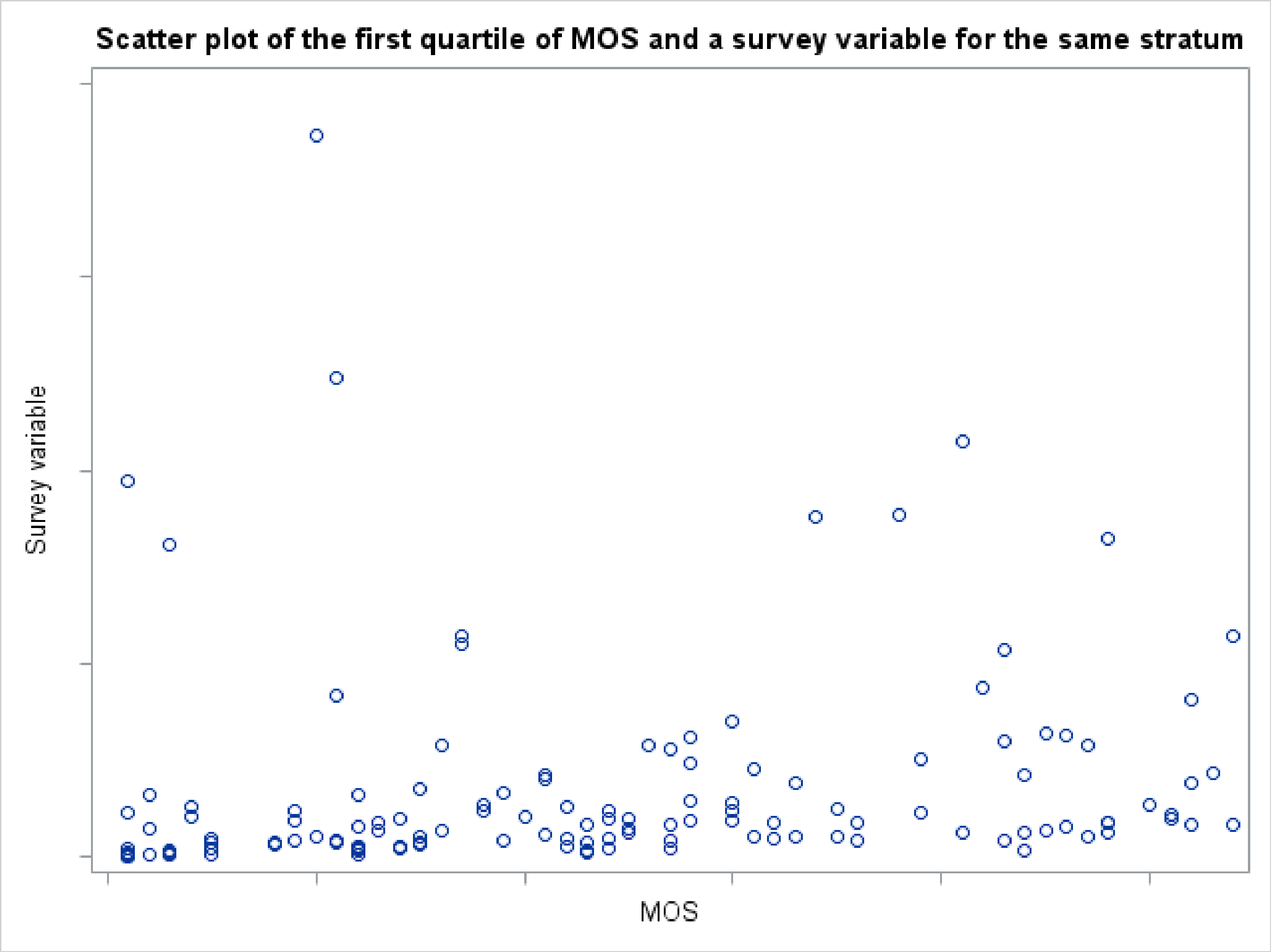
These companies are essentially exchangeable on the survey variable.



United States® Census Bureau

Probability Proportional to Size

Are unequal inclusion probabilities necessary for these companies?



Probability proportional to estimated size

- PPES design will assign unique inclusion probabilities to essentially exchangeable companies
- These inclusion probabilities can be extremely small for some companies
  - Practically excluded from selection
  - Large sample weights $\left(\frac{1}{\pi}\right)$, Max > 670,000!!!
  - Large variance estimates

Winsorization

- We want to apply a one-sided Winsorization procedure to equalize the inclusion probabilities of exchangeable companies.
- Winsorization is a robust estimation technique that “replaces extreme values with less extreme values, effectively moving the original extreme values toward the center of the distribution” (Mulry et. al., 2014)

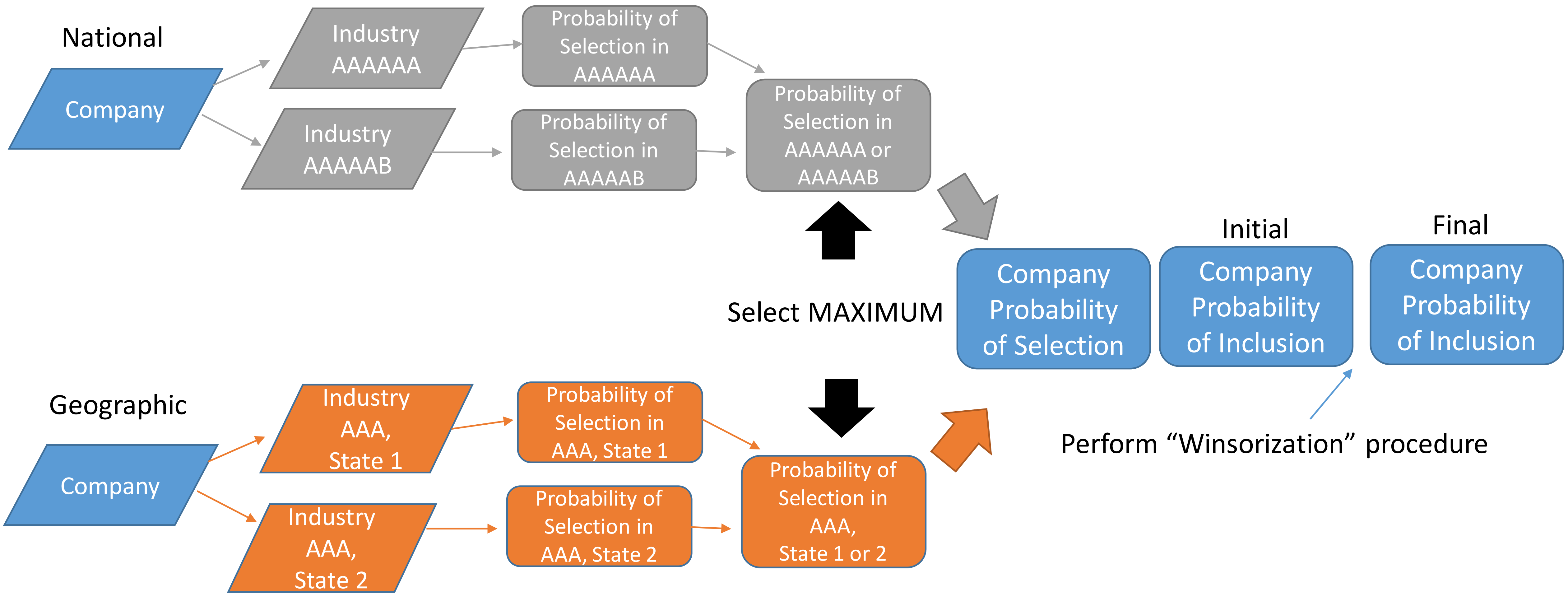
Winsorization of sample weights

- It is common practice for sample weights to be capped at some predetermined threshold **after** sample selection
  - Sample weights $\neq \frac{1}{\pi}$ - > estimation bias
  - Threshold infrequently updated
  - One-size fits all
- The proposed method modifies the sample weights by adjusting sampling probabilities **before** sample selection
  - Sample weights = $\frac{1}{\pi}$
  - Data-dependent thresholds defined for each sampling strata

United States® Census Bureau

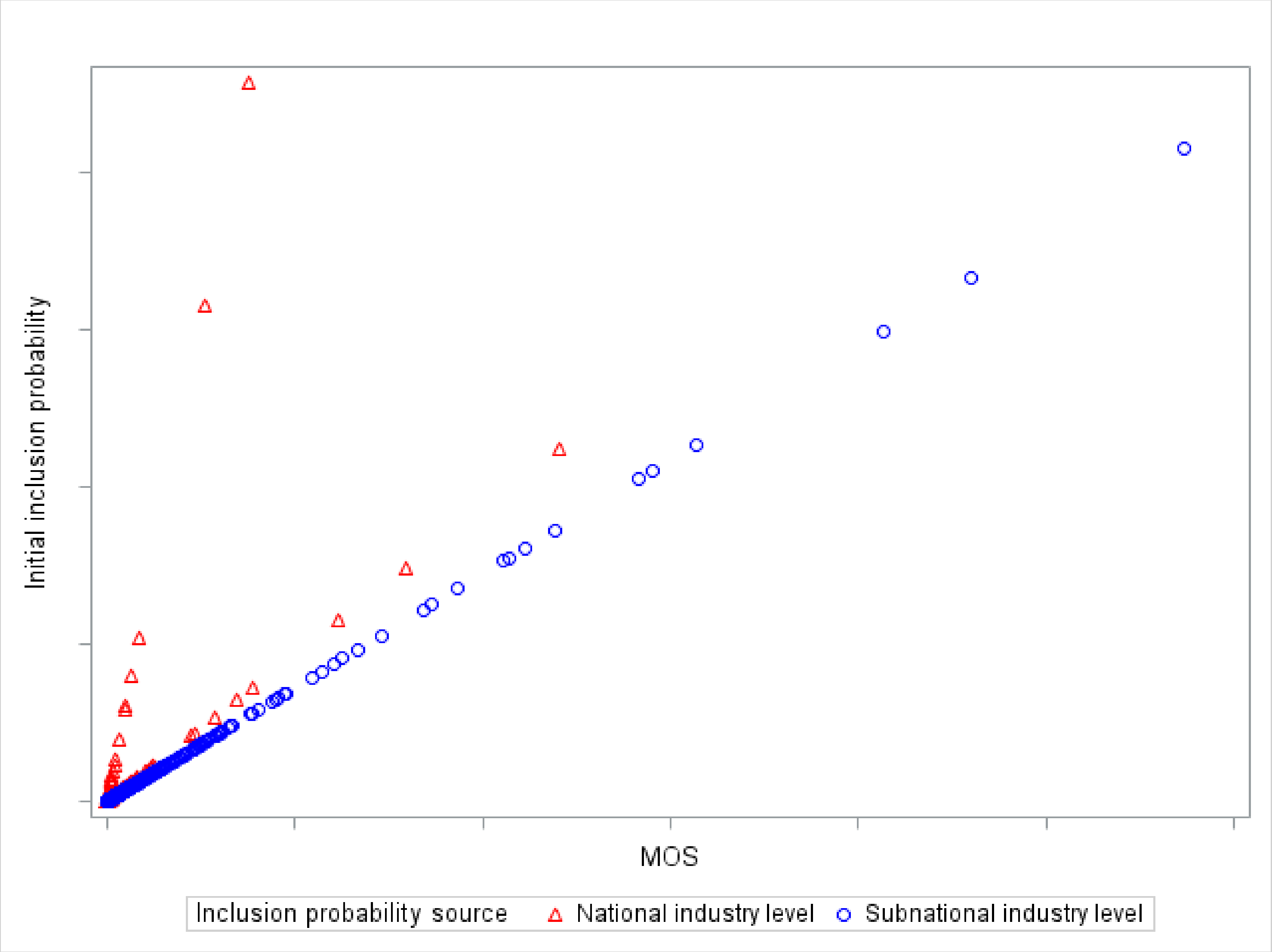
Objectives

1. Maintain unique inclusion probabilities for non-exchangeable units
2. Identify the set of exchangeable units in the lower tail
3. Equalize the inclusion probabilities of exchangeable units (reduce extreme sample weights)



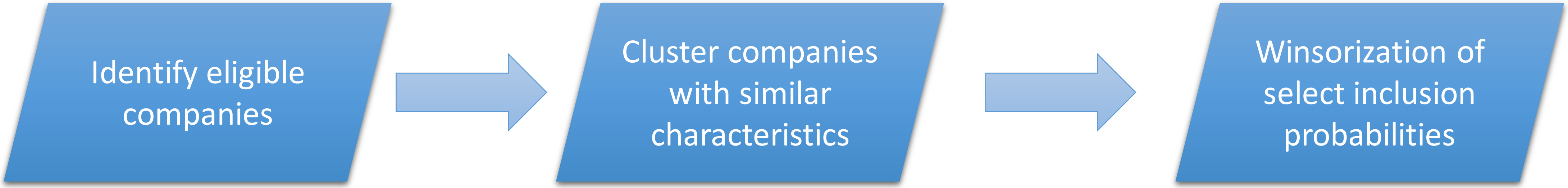
United States Census Bureau

Initial inclusion probabilities by MOS

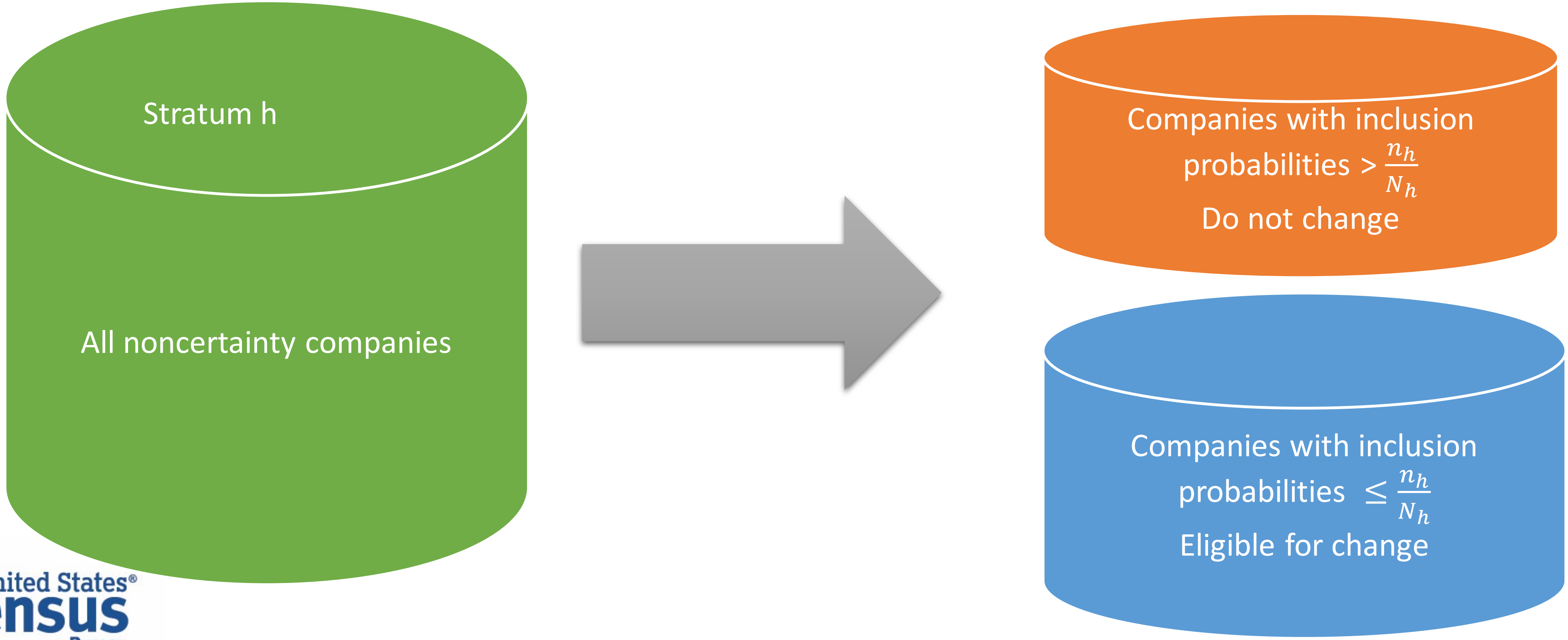


United States® Census Bureau

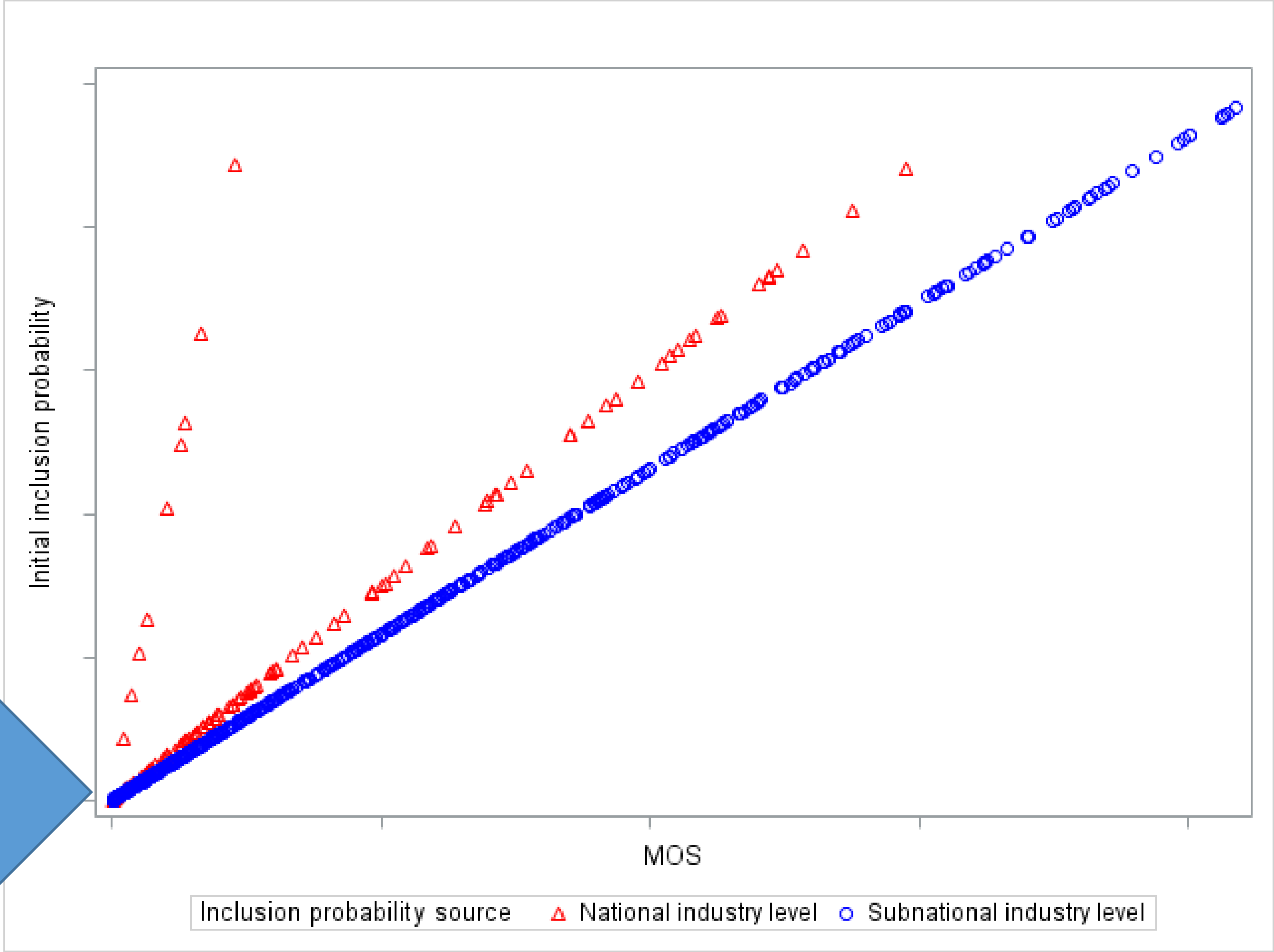
Clustering and Winsorization Procedure



Identify eligible companies



Initial inclusion probabilities by MOS– lower tail



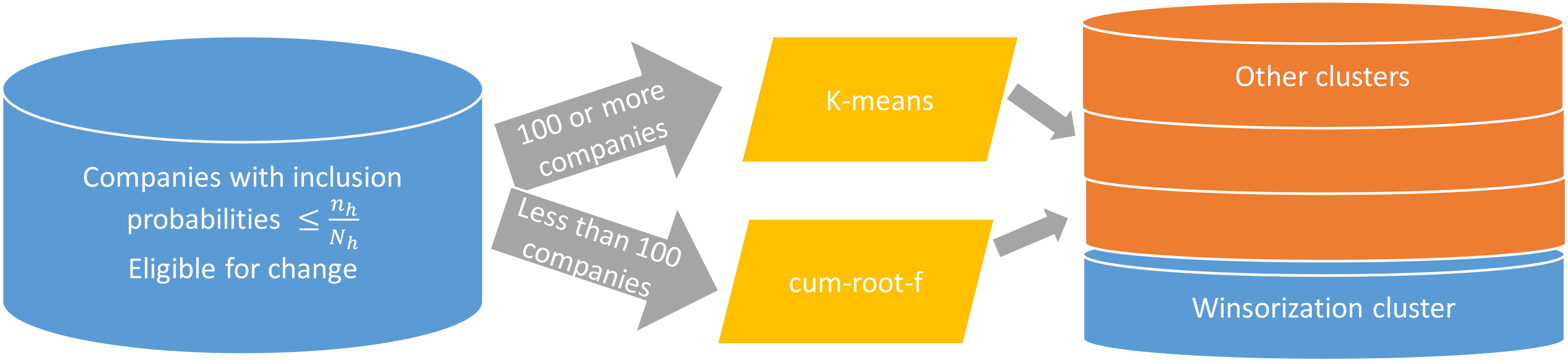
Sample weight = 34,000

Clustering – Two methods

- K-means
  - 100 or more eligible companies
  - Inclusion probability and MOS (standardized first with PROC STANDARD)
  - PROC FASTCLUS
  - K=4 clusters

- Cumulative square root of the frequency (cum-root-f)
  - Dalenius and Hodges Jr, 1959
  - Less than 100 companies
  - Only MOS
  - Number of clusters (2-4) determined by the number of eligible companies in the stratum

Clustering – Two methods



Winsorization cluster

Initial inclusion probabilities

π_{h1}^{NW}	π_{h2}^{NW}	…							π_{hc}^{NW}

Replace inclusion probabilities with the cluster average

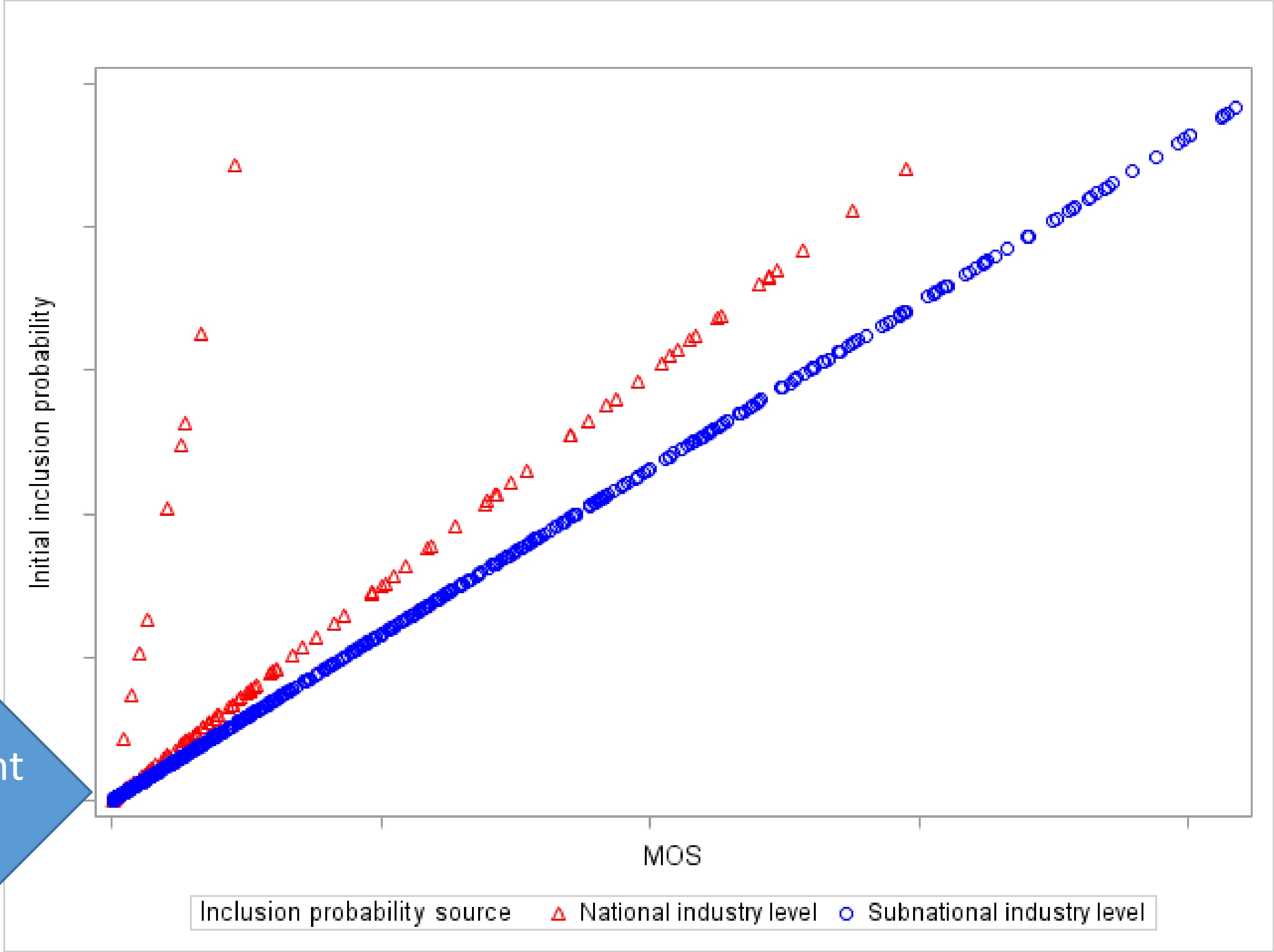
$$\bar{\pi}_h^W = \frac{\sum_{c \in h} \pi_{hc}^{NW} \, I_{hc}^W}{\sum_{c \in h} I_{hc}^W}$$

Final inclusion probabilities

$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$	$\bar{\pi}_h^W$

Initial inclusion probabilities by MOS

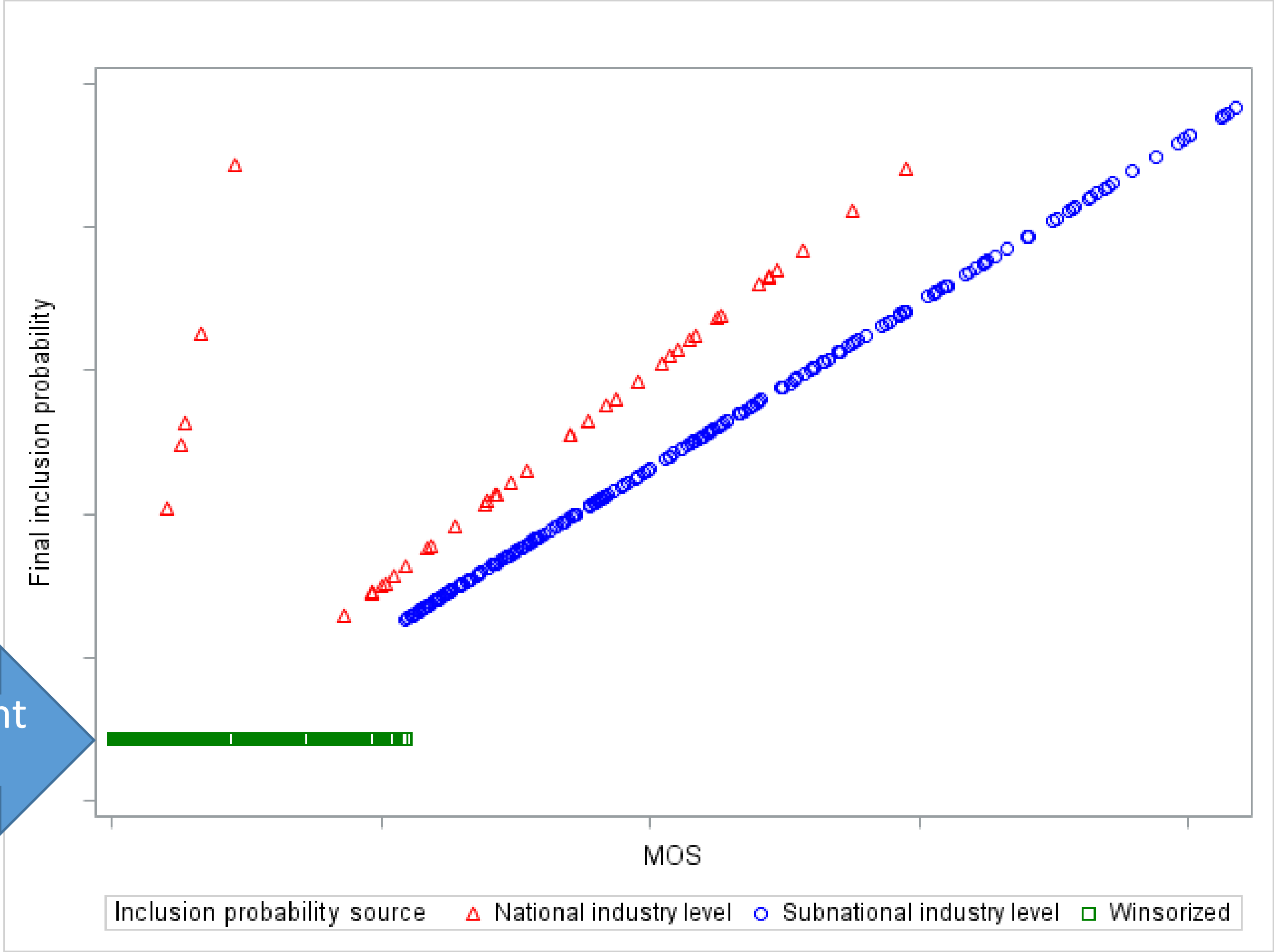
Companies eligible for clustering and Winsorization



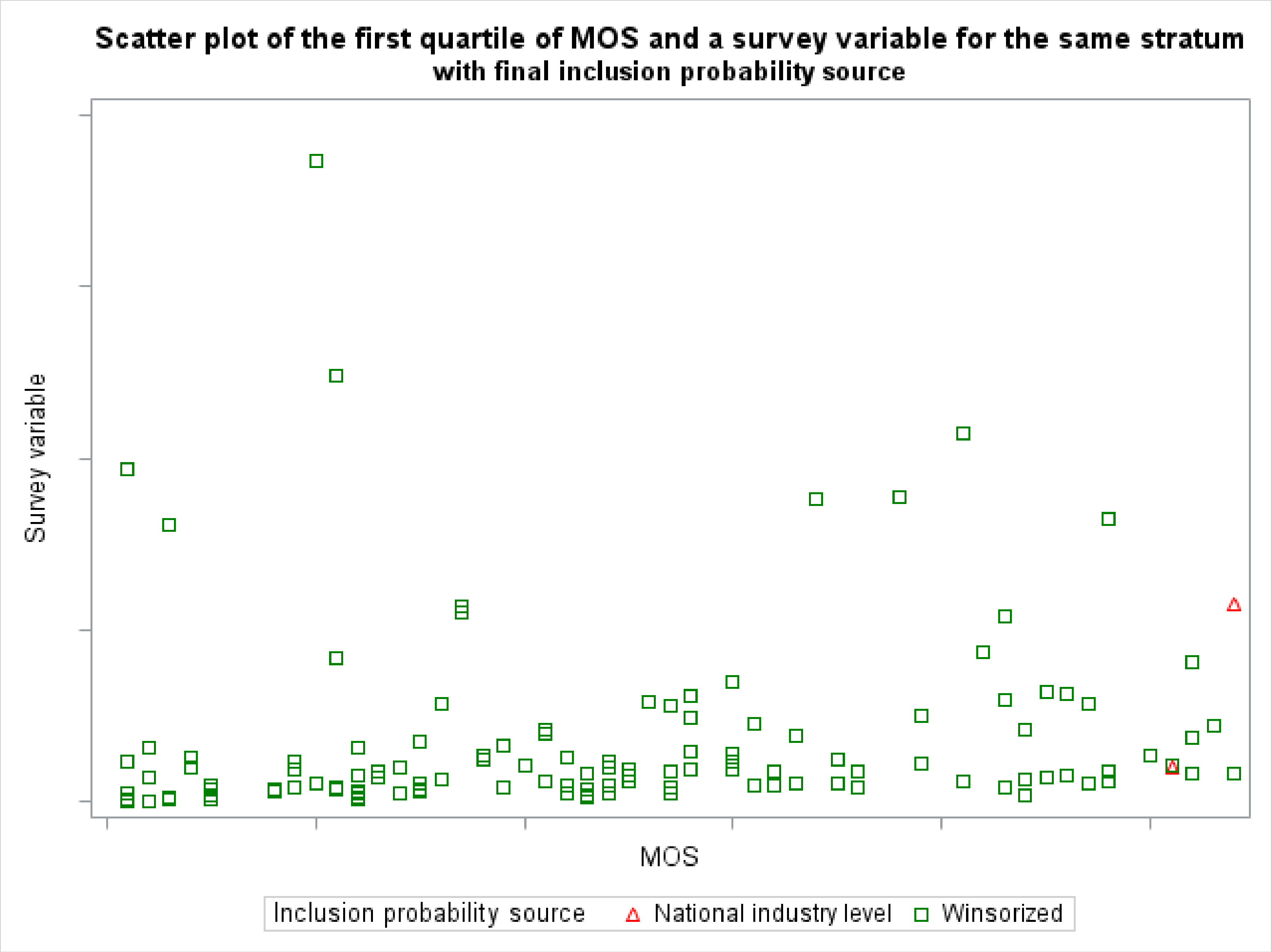
Max sample weight = 34,000

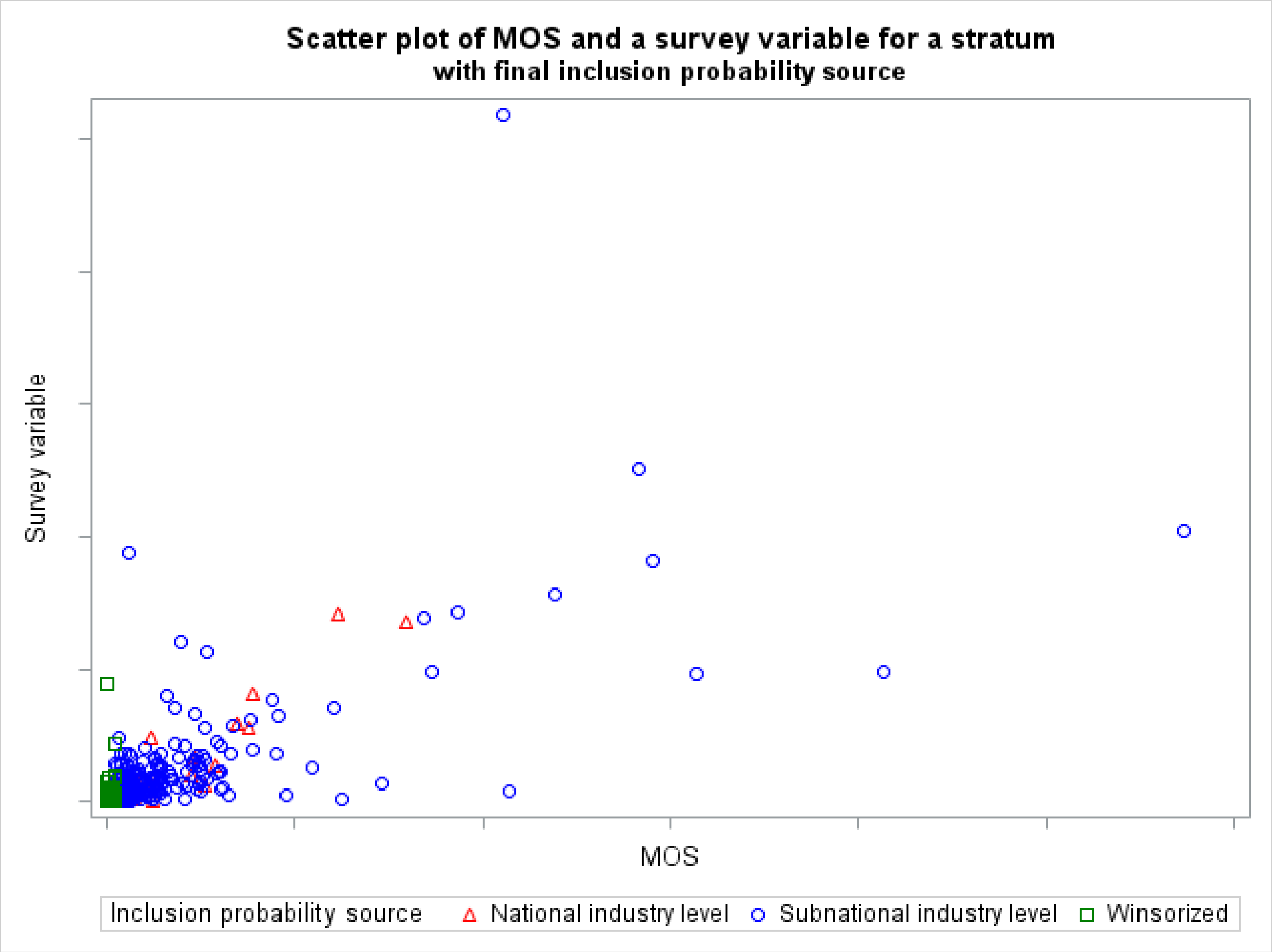
Final inclusion probabilities by MOS

Companies eligible for clustering and Winsorization



Max sample weight = 468

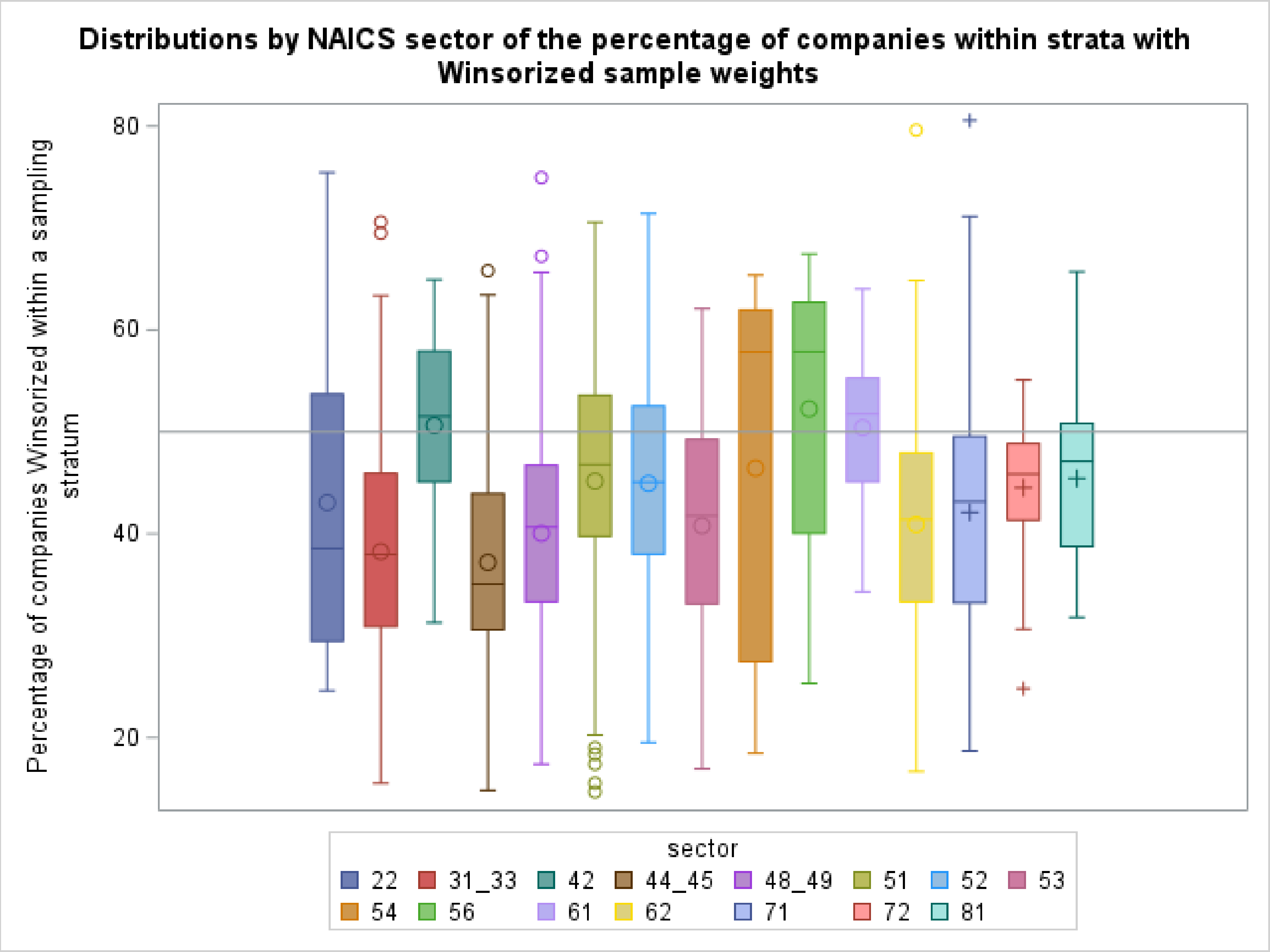

Scatter plot of the first quartile of MOS and a survey variable for the same stratum
with final inclusion probability source
Survey variable
MOS
Inclusion probability source
National industry level
Winsorized

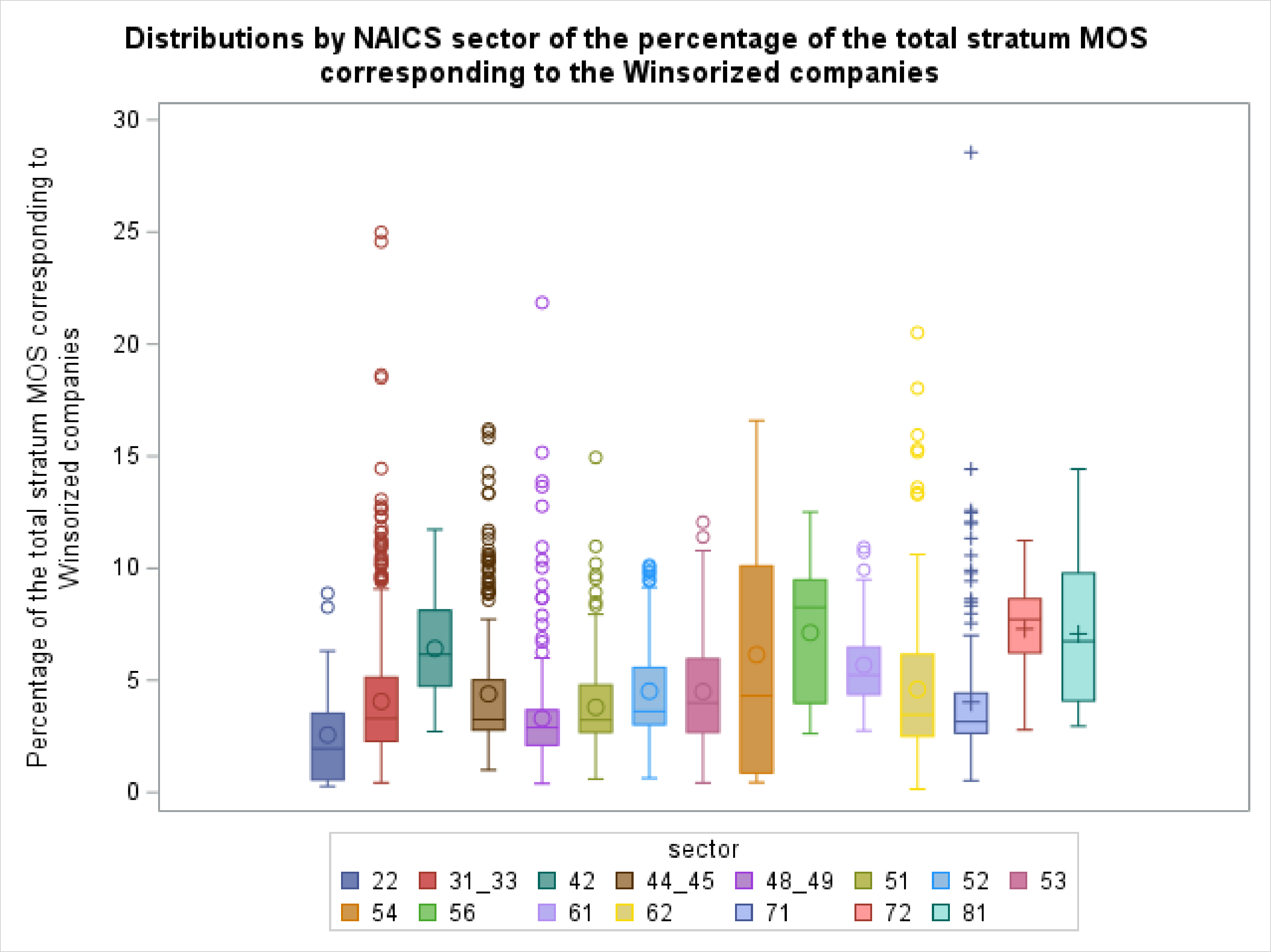

Scatter plot of MOS and a survey variable for a stratum
with final inclusion probability source
Survey variable
MOS
Inclusion probability source
National industry level
Subnational industry level
Winsorized

Results – Sample weights before and after

	Min	Q1	Med	Q3	Max
Winsorized – k-means	13	152	200	260	950
Original Sample Weights	7	128	246	598	598,000

	Min	Q1	Med	Q3	Max
Winsorized – cum-root-f	6	133	272	492	1,410
Original Sample Weights	4	127	319	753	670,000


Distributions by NAICS sector of the percentage of companies within strata with Winsorized sample weights
Percentage of companies Winsorized within a sampling stratum
80
60
40
20
sector
22
31_33
42
44_45
48_49
51
52
53
54
56
61
62
71
72
81



United States® Census Bureau

Summary and future work

- The proposed procedure modifies extreme sample weights by equalizing the inclusion probabilities of exchangeable companies.
- In a way, this marries the equal probability selection of similar sized units from previous surveys with the unequal probability selection of the AIES design and can be used for other PPS designs.
- Future research includes investigating alternative clustering methods for strata with a small number of eligible companies to incorporate the inclusion probabilities into the clustering algorithm and developing a data-dependent method to determine the number of clusters

Acknowledgements

- Katherine Jenny Thompson
- Lucas Streng
- James Hunt
- Andrea Roberson
- Justin Nguyen
- Shelby Plude
- Laura Bechtel

Questions?

Nicole.Czaplicki@census.gov
Yeng.Xiong@census.gov

United States Census Bureau