Foxes in Alaska

New Analytics Methods in Social Media: Population Estimation Informed by Wildlife Biology

Speaker:
Prof. Dr. Dr. Dietmar Janetzko **Cologne Business School** **@dietmarja**

49TH FIBEP WORLD MEDIA INTELLIGENCE CONGRESS
BERLIN 4-6 October 2017

@_FIBEP
#FIBEP
#WMIC17

CBS COLOGNE BUSINESS SCHOOL

Table of Contents

- Capture Mark Recapture (CMR)
- Applying CMR to Twitter Analytics
- Real World Example – Tweets mentioning “Trump”
- Discussion

Population Estimation via Capture Mark Recapture (CMR)

Capture

↓

Mark

↓

Recapture

↓

Possibly repeated Recaptures

1.) 1st Sample
Random Selection of Animals, e.g., Foxes

2.) Animals captured are marked so they can be identified

3.) 2nd Sample
2nd random selection of animals, marked animals are counted (recaptured=previously marked

*n*th Sample

Capture-Mark-Recapture (CMR) Models

Peterson (1886) Peterson-Lincoln model: The most basic

Lincoln (1930) capture-mark-recapture (CMR) model

Schnabel (1938) Generalization of the Petersen-Lincoln method to multiple samples

Cormack(1964)

Jolly(1965) Generalization to open populations

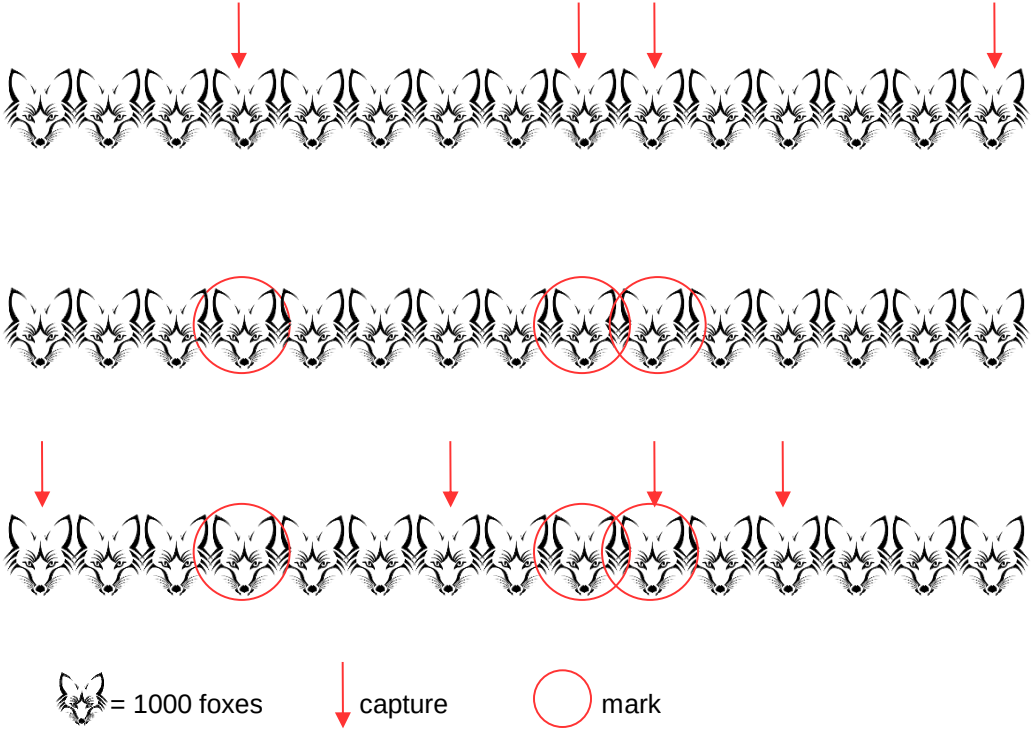
Seber(1965)

From "Foxes in Alaska" to Tweets

	CMR (Animals)	CMR (Twitter)
Individuals	Animal	Tweet
Catching	Catching animals	Observing Tweets
Population	Total number of animals of a species of interest at time t	Total number of Tweets of a category of interest at time t
Open Population	Population affected by immigration and emigration	Twitterers joining in or leaving Twitter

Population Estimation Capture-Mark-Recapture

Say, the number of Foxes in the Population of Interest = 15.000 (usually not known!)



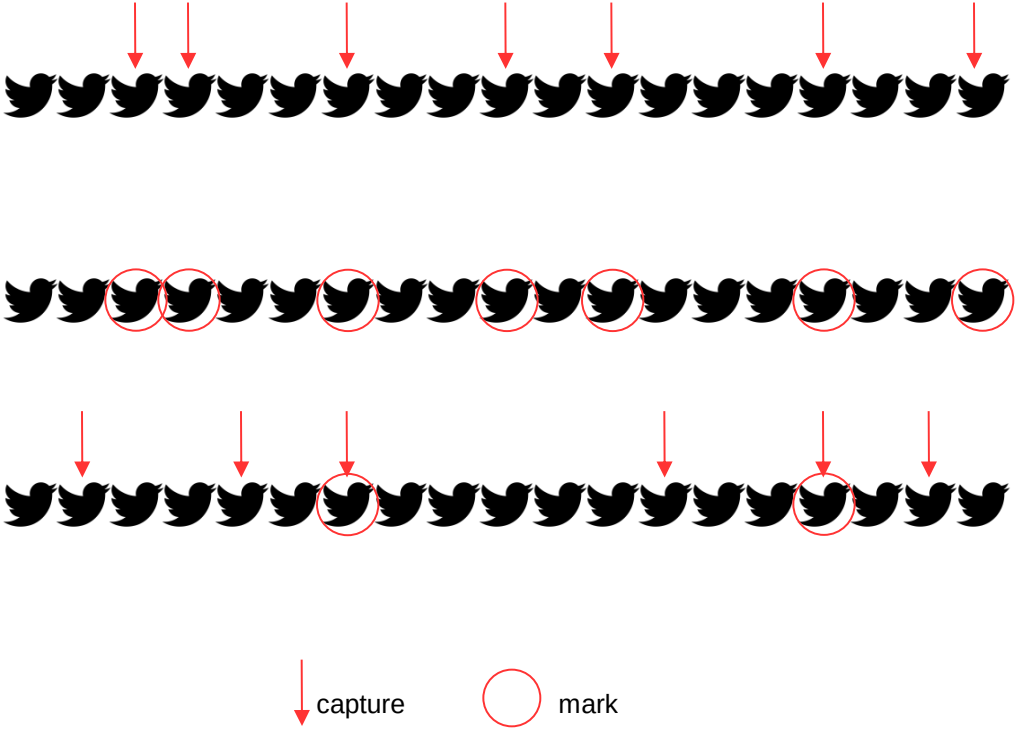
1.) Capture / 1st Sample
Random Capture of 4000 Foxes

2.) Mark
The foxes are marked and released again

3.) Recapture / 2nd Sample
Random selection of 4000 foxes
1000 of which found to carry a mark

Population Estimation via Petersen-Lincoln Index

Total number of Tweets in the Population of Interest = 19 (usually not known!)



1.) 1st Sample (Capture)
Random Selection of 7 Tweets via Twitter's free Streaming API (1 %)

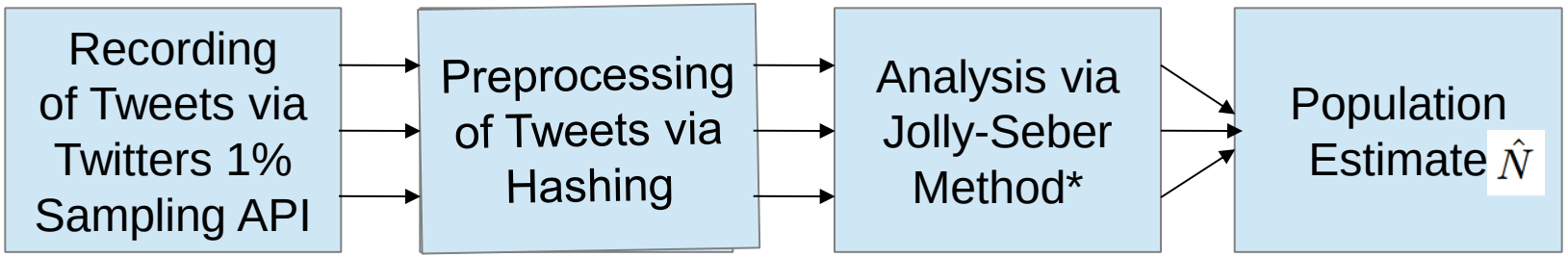
2.) Tweets are recorded (Mark)

3.) Sample (Recapture)
Random Selection of 6 Tweets 2 of of which known from the 1st sample

How to Apply CMR to Twitter Analytics:

5-day Collection of Tweets that mention “Trump”

via 3 independent Apps (start date Oct 1st 2017)



* Facilitates CMR analysis in *open populations*, i.e., populations affected by death birth, emigration, immigration.

Recoding Tweets via a Deterministic Cryptographic Hash Function

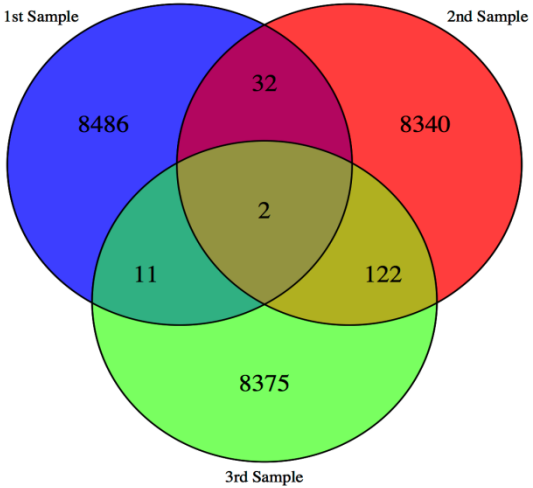
“33158704 @AP: Chancellor Angela Merkel bids for fourth term as Germans head to the polls. https://t.co/PRm8ZeGszG”

Twitter User ID

Tweet

"**b16827bf9535f4c61d69351bfe4d73e3**" (fixed-sized bit string)

Collection of Tweets that mention “Trump” via 3 independent Apps – Results after 24 hours



n_1 Size of the 1st sample = 8531 Tweets *

n_2 Size of the 2nd sample = 8496 Tweets *

n_3 Size of the 3rd sample = 8510Tweets *

n_{12} Number of Tweets recaptured (overlap) = 35

n_{123} Number of Tweets recaptured (overlap) = 135

~~Estimated Population (Tweets with the keyword “Trump” in the time-frame of 24 h considered, Jolly-Seber method for open populations)~~

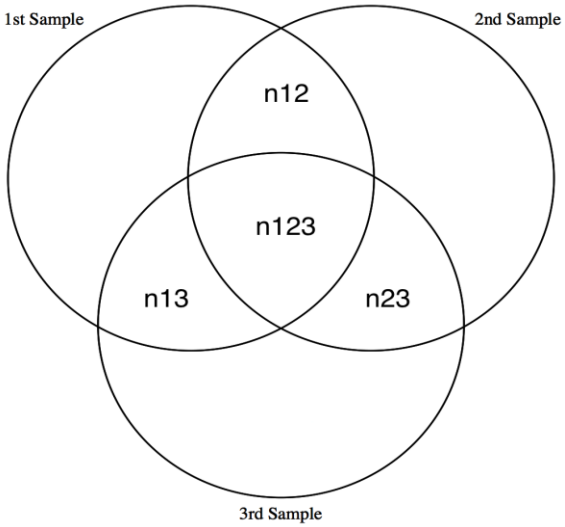
= 189 783

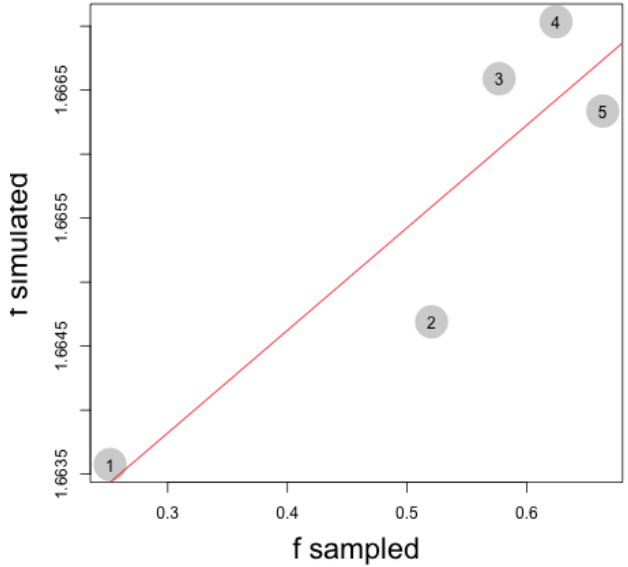
$\hat{N}$

* Only unique Tweets (unique Twitter ID, unique Text) are counted, Repetitions *within* M or n are not considered here

Results : Collection of Tweets that mention “Trump” via 3 independent Apps

Are the population estimates correct? To answer this question a first preliminary simulation study has been conducted. The simulation used 3 sets of random hash strings with the same size as the sets obtain from Tweet collection. The evaluation metric used for both the simulated data and the Tweet data was $f = n_{12}+n_{23}+n_{13} / n_{123}$





Discussion

- Rests on word-identical Tweets – adaptations are needed when Tweets are varied
- Estimating population-size on the basis of possibly huge amounts of data
- In CMR models, not only population estimators are available but also estimators for survival and mortality.
- Further evaluations with Twitter data (Firehose) are necessary

Thank you!