False Positive Reduction: An Algorithmic Approach

False Positive Reduction: An Algorithmic Approach

K. Nebiolo

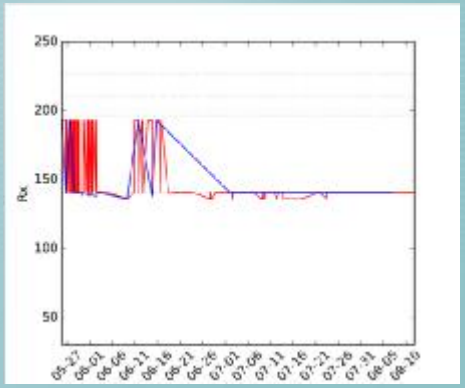
T. Castro-Santos

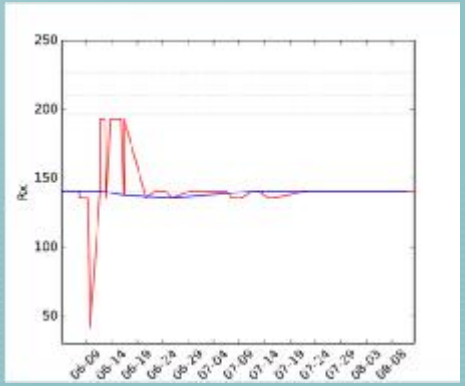
Kleinschmidt

USGS
science for a changing world

What are false positives? Why do we care?

- Problems with telemetry data of all types (Beeman & Perry 2012)
  - Received signal may not correctly assigned
  - Not all records generated by receivers are from tagged fish
  - Not all tagged fish are recorded when present
- This introduces false positives into our data
  - Bias results in favor of presence and overestimate frequency of occurrence
  - Or overestimate residence time within critical infrastructure, which increases measures of delay





Current False Positive Reduction Methods

- Current false positive reduction methods rely simple metrics and subjective opinion
  - Power floors
  - Consecutive detections (2, 4, 6, etc…) (Beeman & Perry, 2012)
  - Logical errors in site progression
- Manual classifications are labor intensive
  - For very large studies with many releases and sites (whole river studies) classification can become cost prohibitive

Naïve Bayes Classifier (Minsky 1961)

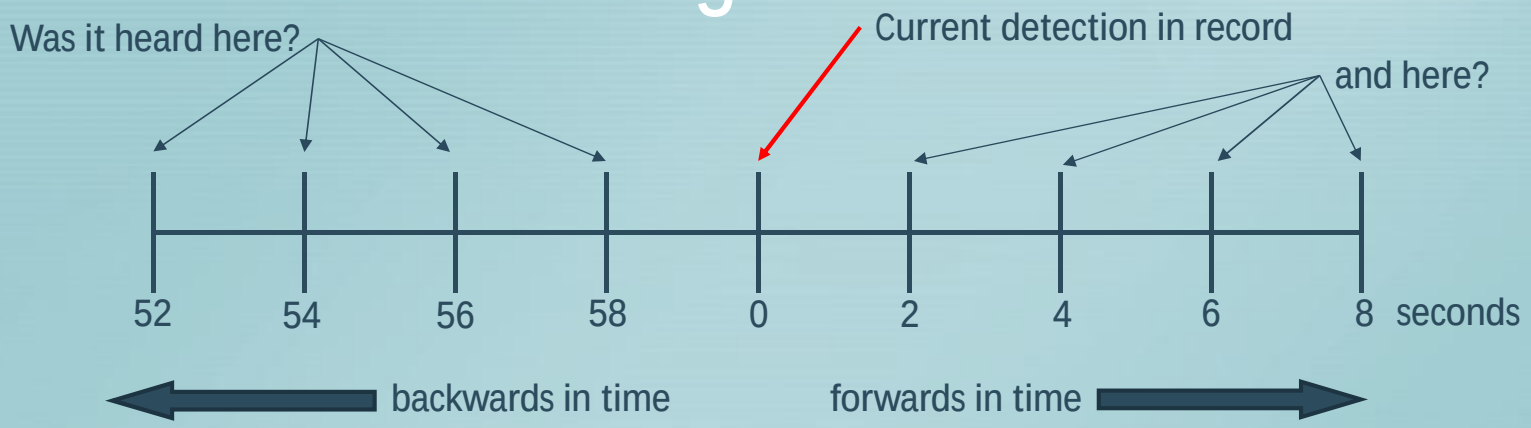
- Bayes rule (conditional probability) can estimate the probability that a record is either true or false positive given observed data
- *Prior*: marginal probability for a detection class (C)
- *Likelihood*: The conditional probability of an observable event ($F_1, \ldots, F_n$) given each state of nature (C)
- *Posterior*: the probability of the class occurring given the observable event

Posterior Prior Likelihood

$$P(C|F_1, \ldots, F_n) \propto P(C) * \prod_{i=1}^{n} P(F_i|C)$$

- Training data set contains observations of feature variables ($F_i \ldots F_n$) and known detection classifications (C)
- Naïve assumption

Predictors – Creating the Detection History



P	X	P	X	P	X	P	P	P
1	0	1	0	1	0	1	1	1

Predictors – continued

- Derived from detection history
  - Hit Ratio
  - Consecutive Hit Length
- Power
- Noise Ratio

Detection History (+/- 4)									Hit Ratio	Consecutive Hit Length
1	0	1	0	1	0	1	1	1	6/9	3
1	0	1	0	1	0	1	0	1	5/9	1
1	1	1	1	1	0	0	0	0	5/9	5
1	1	1	1	1	1	1	1	1	9/9	9

Implementation

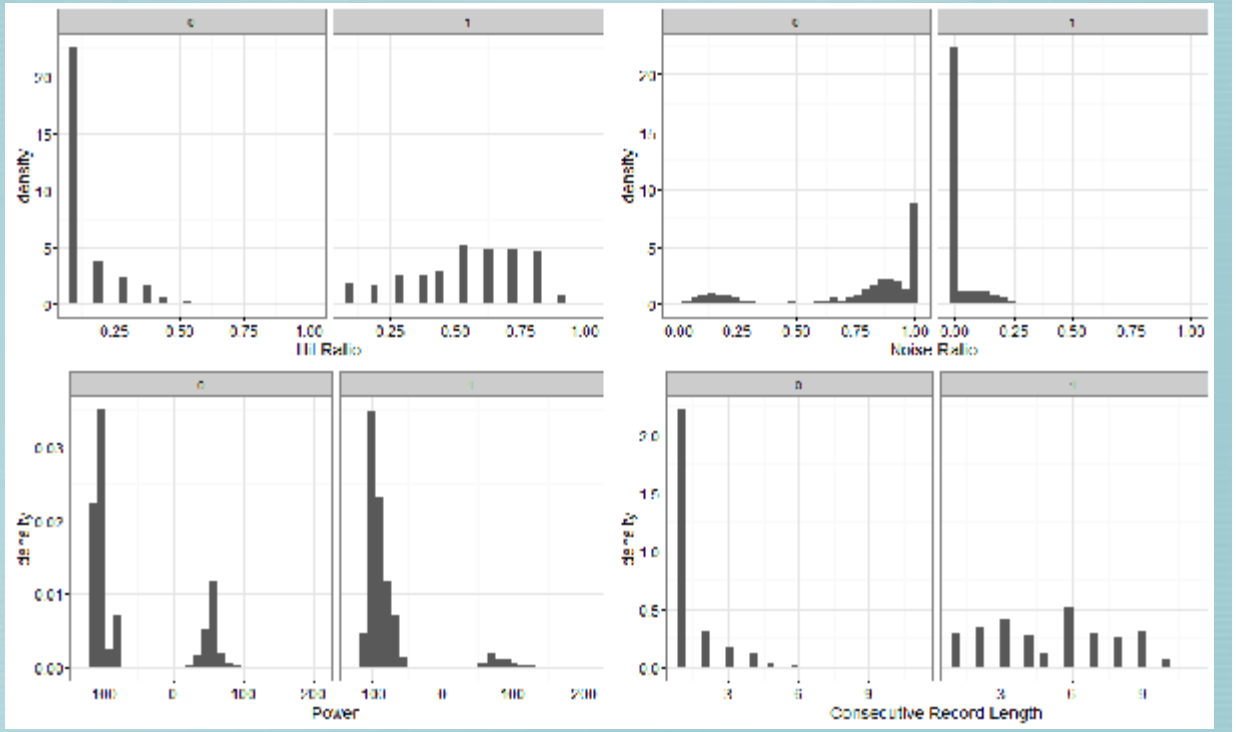
- Algorithm split into two parts, *Training* and *Classification*
- Training:
  - Beacon tags placed at strategic locations throughout the study area - provide information on what known true positive detections look like
  - Miss coded and noise transmissions provide information on what false positive detections look like
  - Loop over all beacon tag and miss coded detections and derive metrics
    - (detection history, hit ratio, etc.)
  - Store to RDBMS in this case SQLite
- Classification:
  - Loop through records for known study tags,
  - Derive metrics
  - Calculate posterior for true and false positive – use MAP to classify
- Algorithm Accuracy assessed with k-fold cross validation (k = 10)
  - Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

Beacon Tag Locations

Results – Orion Training

- In total – 2,644,990 beacon tag hits and 517,881 known false positives

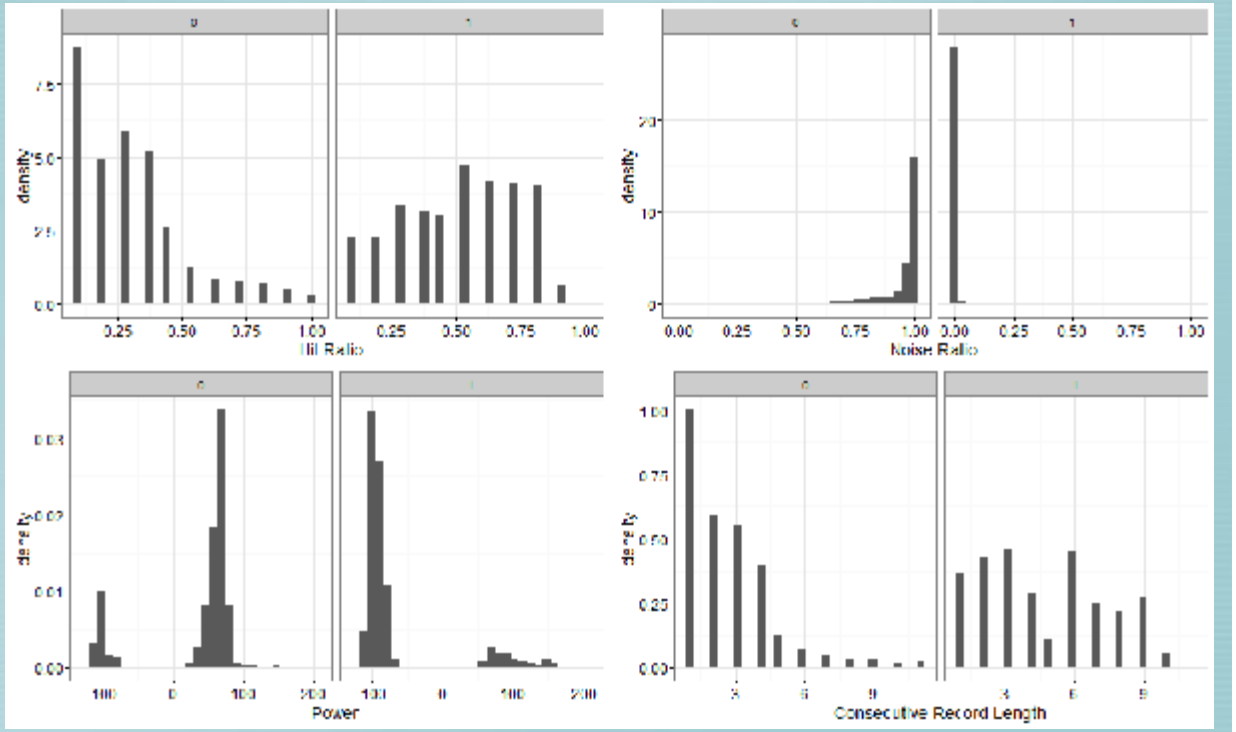
	Classified		
	False	True	
True Negative	479,349	38,532	NPV: 0.93
True Positive	39,745	2,605,245	PPV: 0.98
	Specificity: 0.92	Sensitivity: 0.99	



Results – Lotek Training

- In total – 2,215,818 beacon tag hits and 331,612 known false positives

	Classified		
	False	True	
True Negative	298,927	32,685	NPV: 0.90
True Positive	5,073	2,210,745	PPV: 0.99
	Specificity: 0.98	Sensitivity: 0.99	

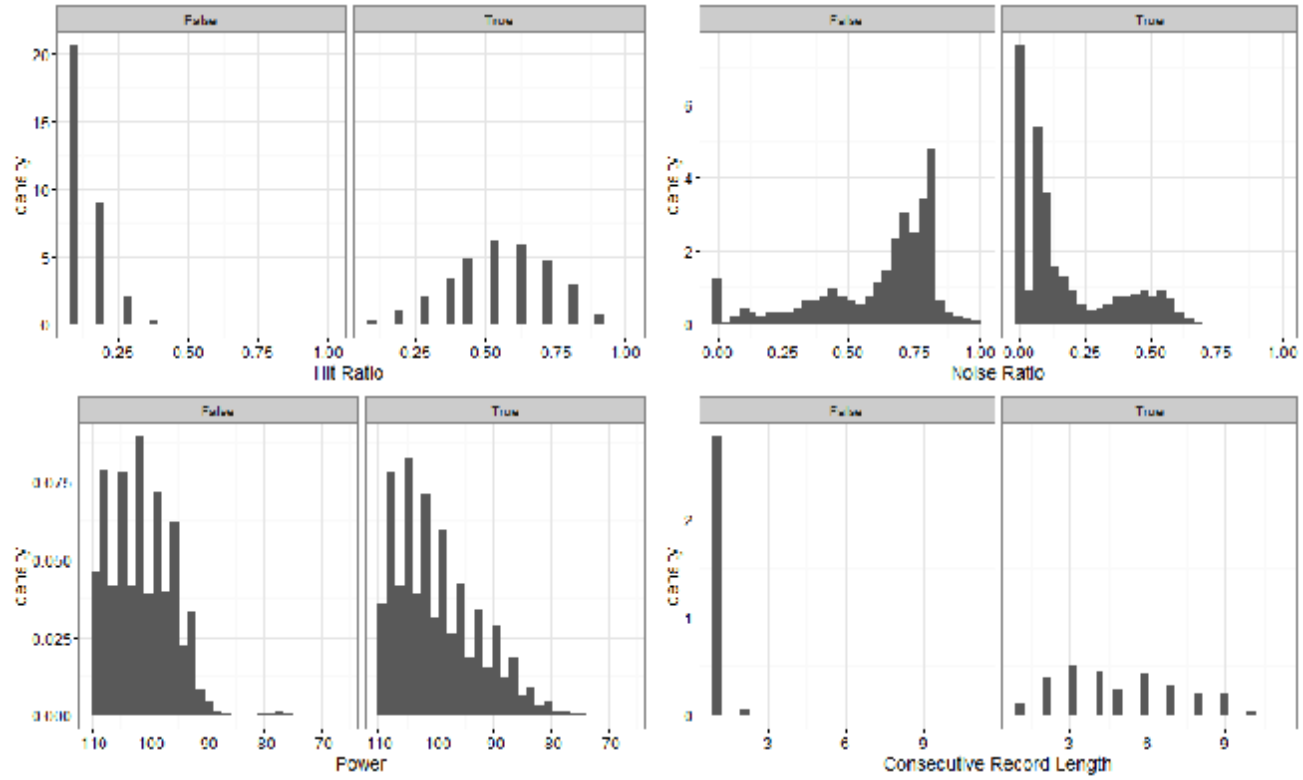


Results – Orion Classification

575,548 study tag detections

- 503,347 true
- 72,201 false positive

14% False Positive removal rate

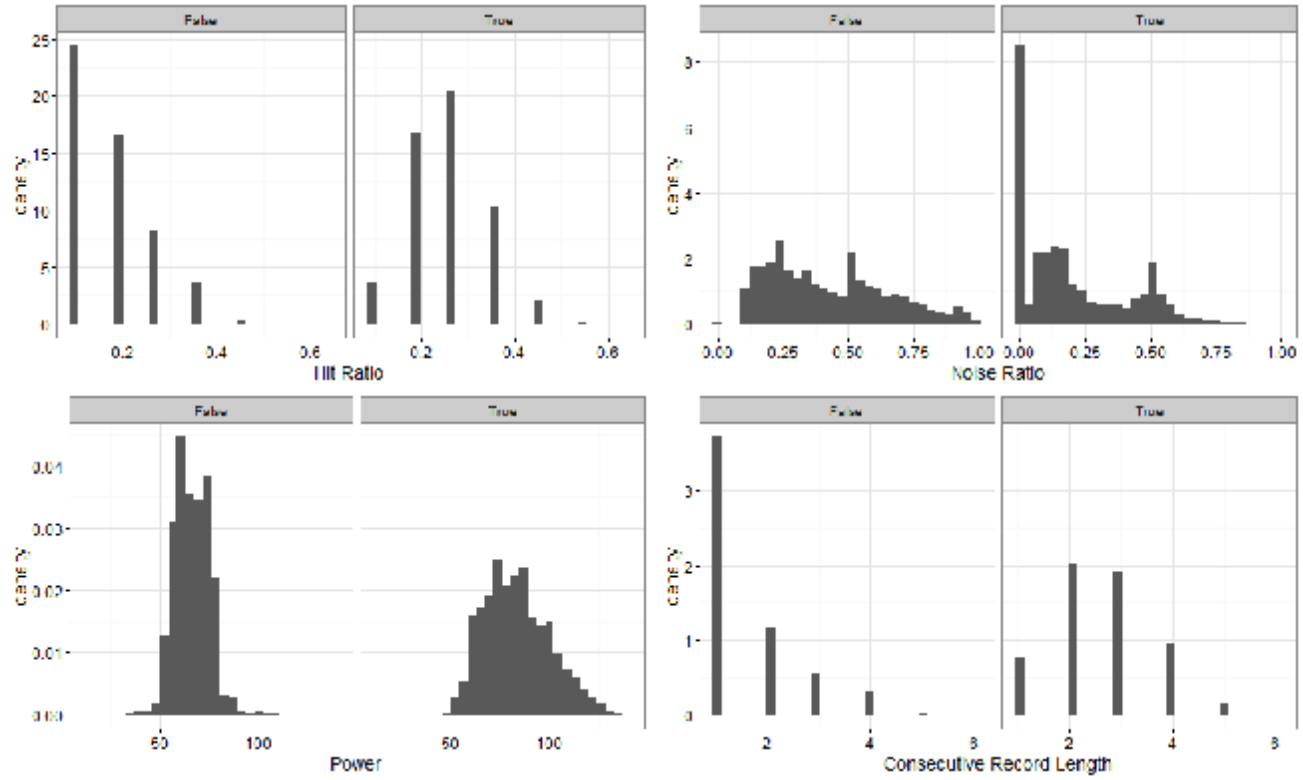


Results – Lotek Classification

131,088 study tag detections

- 106,693 true
- 24,395 false positive

23% False Positive removal rate

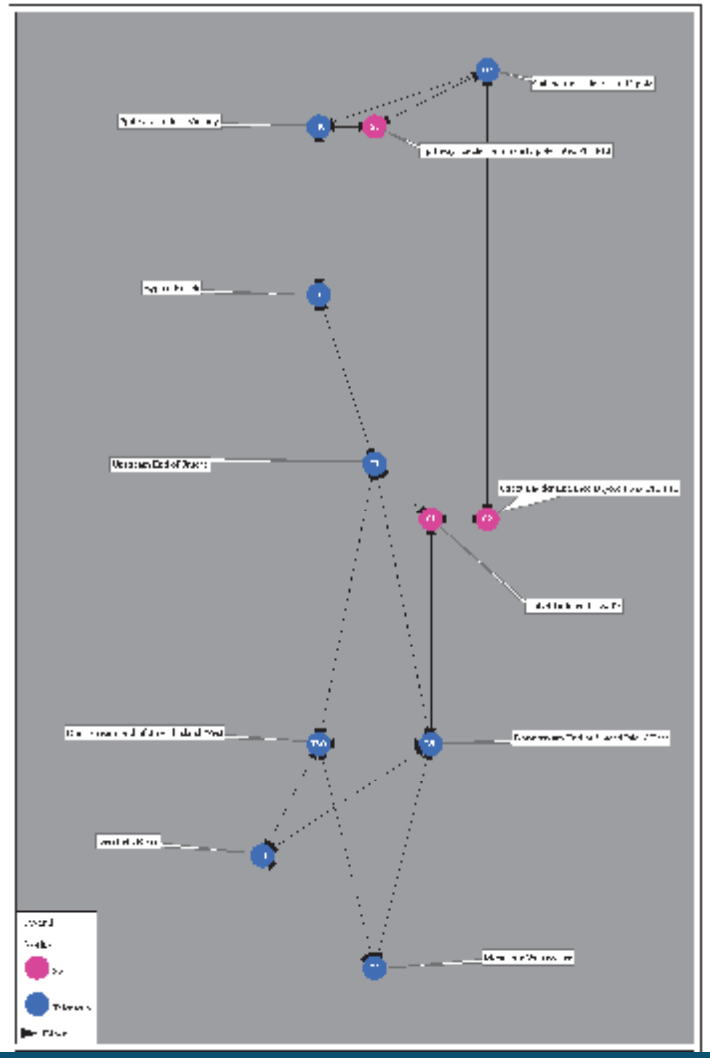


Conclusion

- Once the algorithm is up and running it takes minimal supervision
  - ~ 15 minutes every hour
  - Computationally, detection history creation most time consuming part
- Does a great job of removing garbage detections with minimal effort
- Does not do a good job at discriminating position when receiver detection zones overlap
- From raw data through statistical analysis with 300 tagged at 14 stations – <u>2 weeks</u>

Future Research

- Investigate lag time between detections
  - Good detections = stable lag
  - Bad detections = varying lag
- Utilize network/graph topology and investigate ways to pinpoint fish in space & time
  - 2nd round data classification
- Make training data better



Beeman, John W. and Russel W. Perry. "Telemetry Techniques: A User Guide for Fisheries Research." Ed. N.S. Adams, J.W. Beeman and J.H. Eiler. American Fisheries Society, 2012. 505-517.

Works Cited

Beeman, John W. and Russel W. Perry. "Telemetry Techniques: A User Guide for Fisheries Research; Chapter 9: Bias from False-Positive Detections and Strategies for their Removal in Studies Using Telemetry." American Fisheries Society, 2012. 505-517.

Minsky, Marvin. "Steps toward artificial intelligence." *Proceedings of the IRE* 49.1 (1961): 8-30.

Thank you

Kevin Nebiolo, PhD.
Kevin.Nebiolo@kleinschmidtgroup.com

Ted Castro-Santos, PhD.