

Theory and practice of model-based data integration for modeling species' distributions

Saras Windecker & David Uribe
Species on the Move 2023

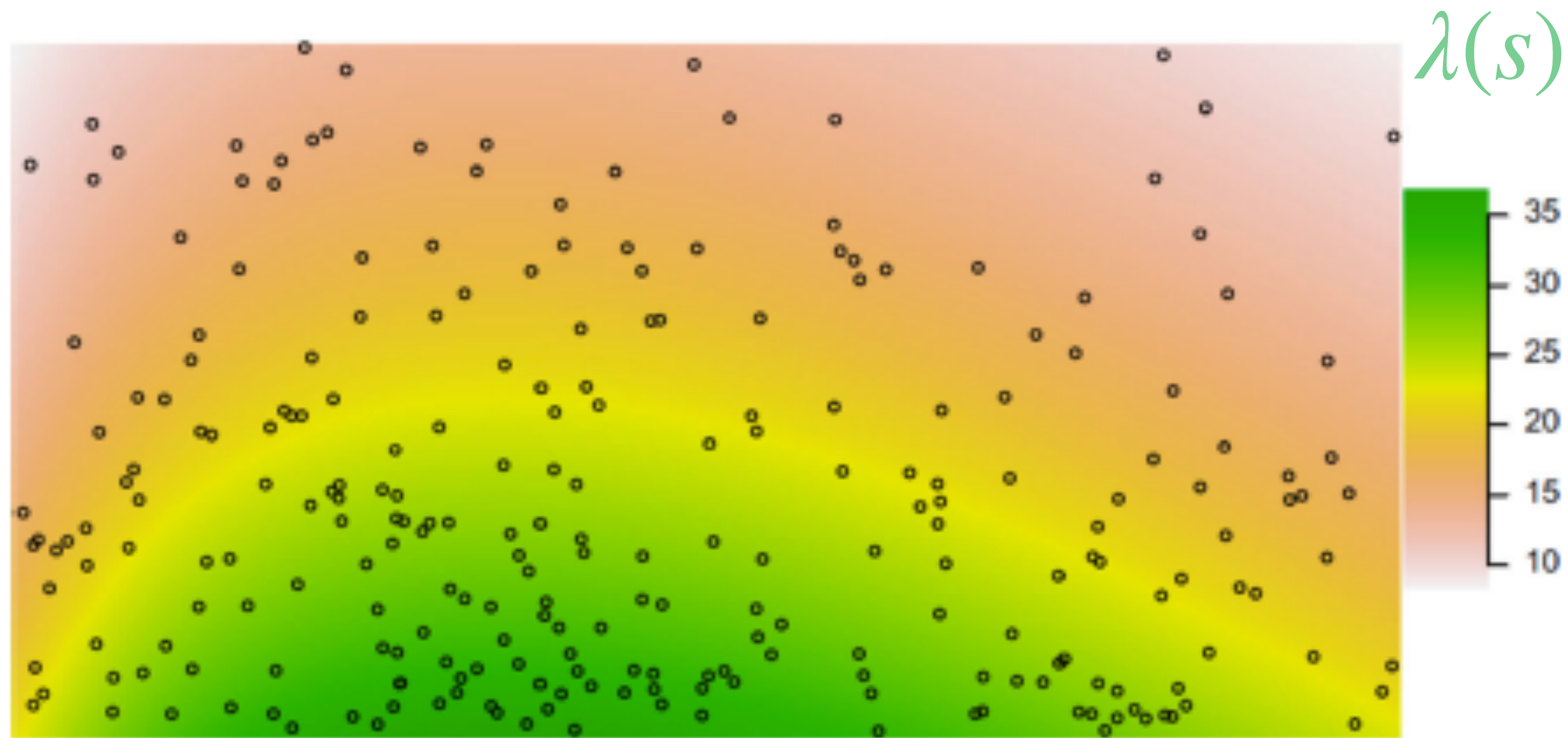
**How did we come to
this?**



What is the aim?

What is the aim?

Distribution of individuals of a species (relative abundance) across space.



Fithian W, Elith J, Hastie T, Keith DA. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol Evol.* 2015;6(4):424-438. doi:10.1111/2041-210X.12242

Data integration combines datasets by explicitly modelling their data collection processes, incorporating their biases, and propagating as much information as possible about the process of interest.

What is the motivation for data integration?

What is the motivation for data integration?

Capitalise on many different data types, none of which are perfect.

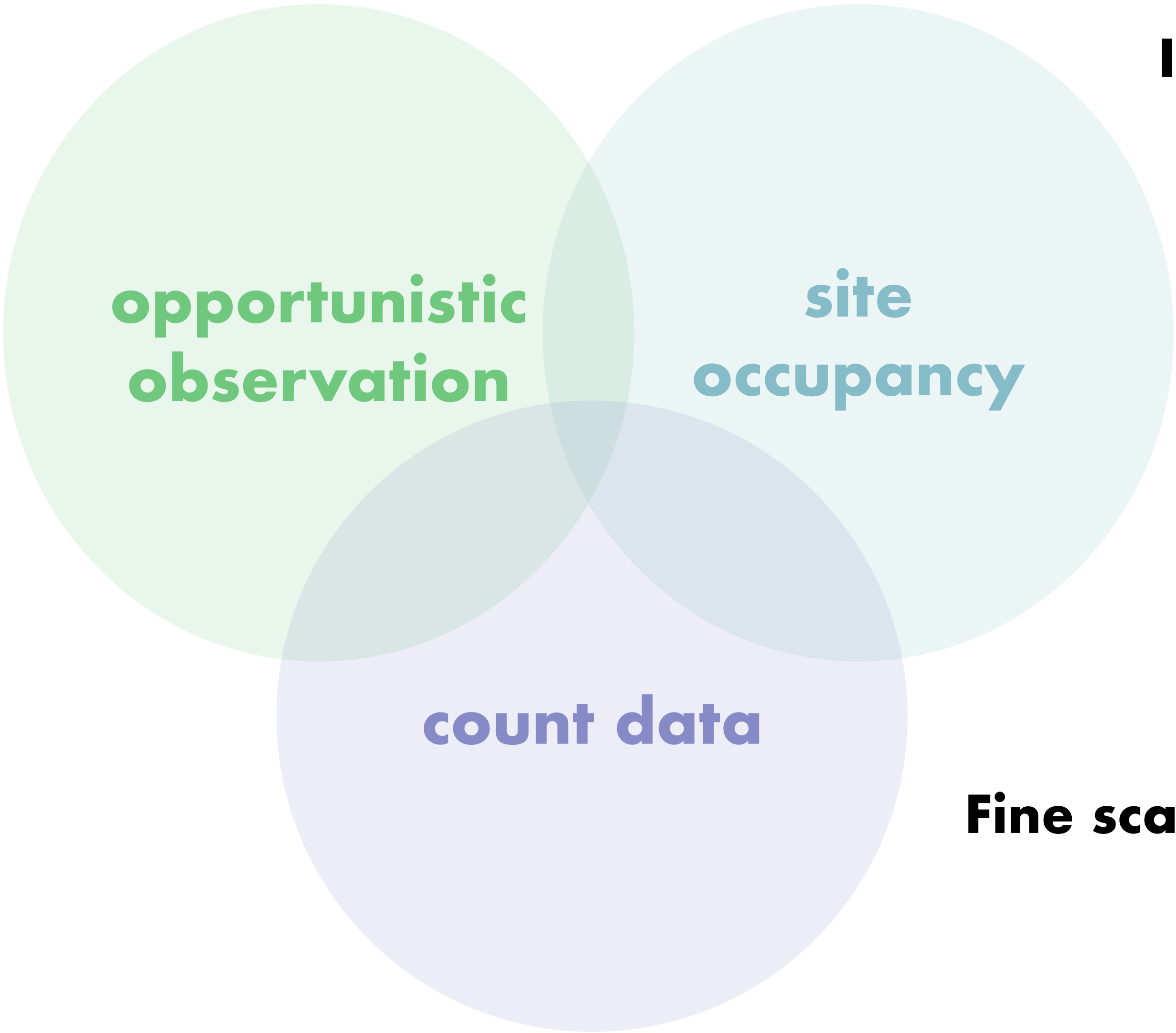


Vary in extent.

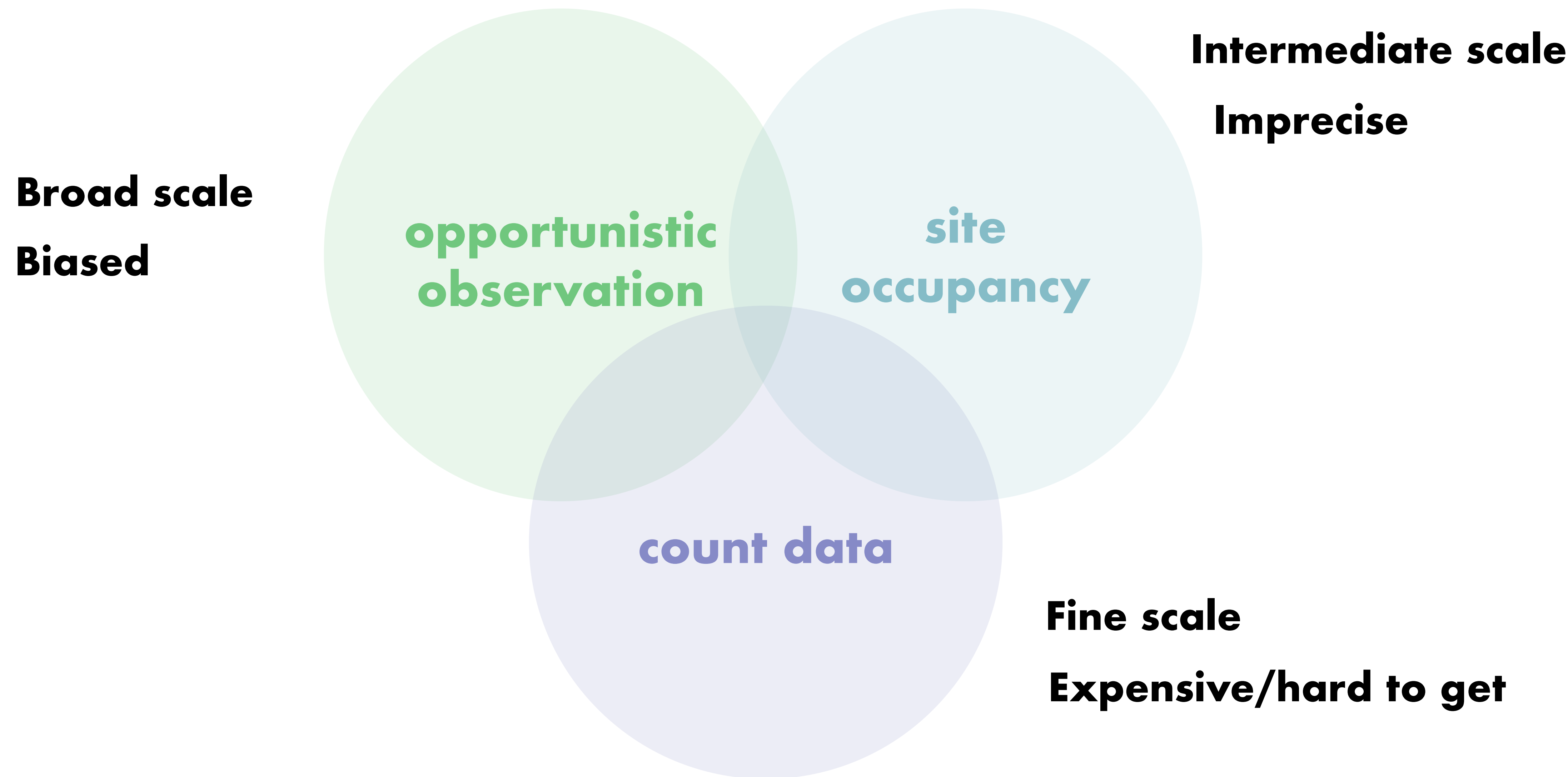
Broad scale

Intermediate scale

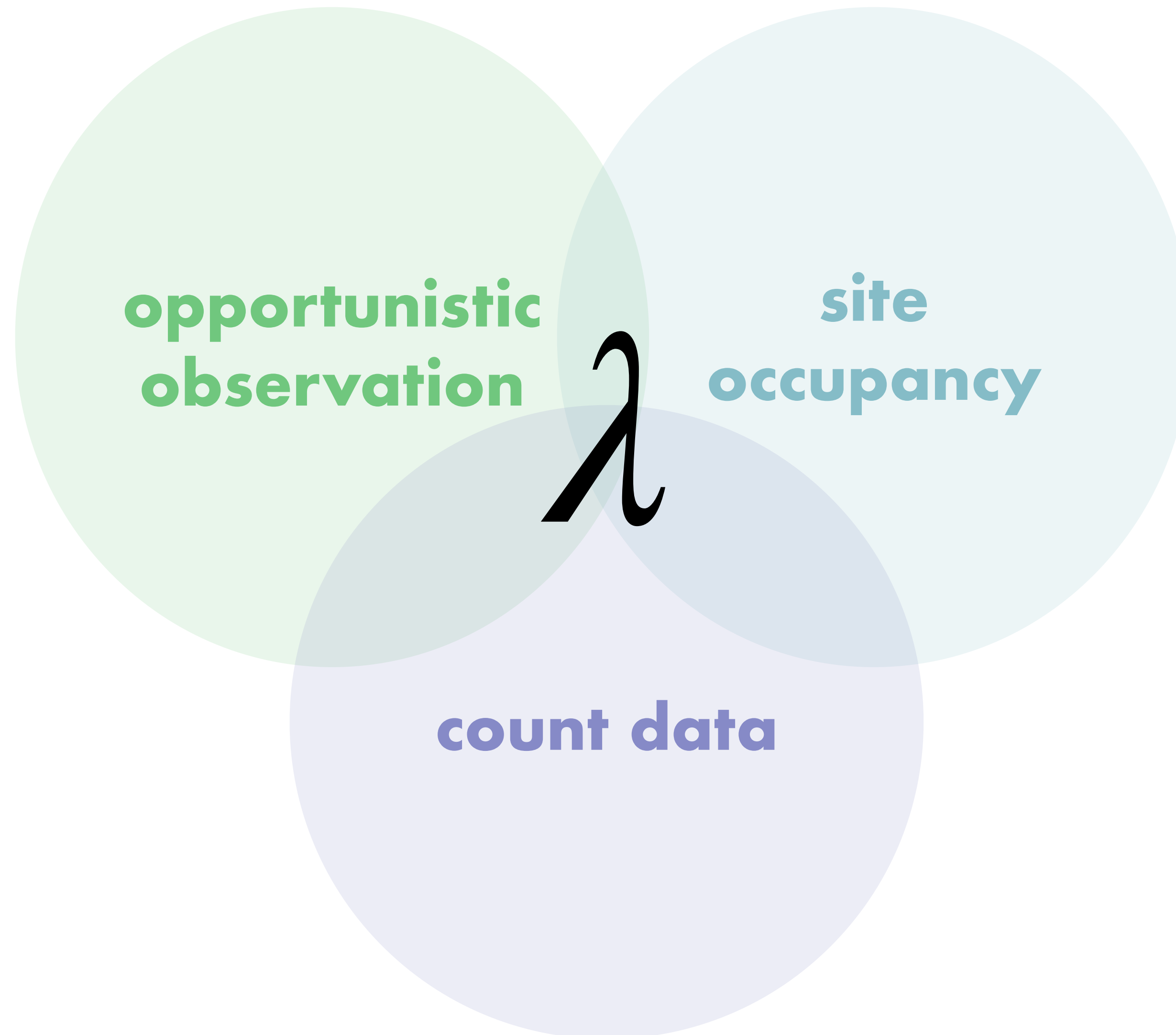
Fine scale



Vary in extent. And unique biases.



Can use a mechanistic link between data types to establish a common parameter of spatial distribution.



count data

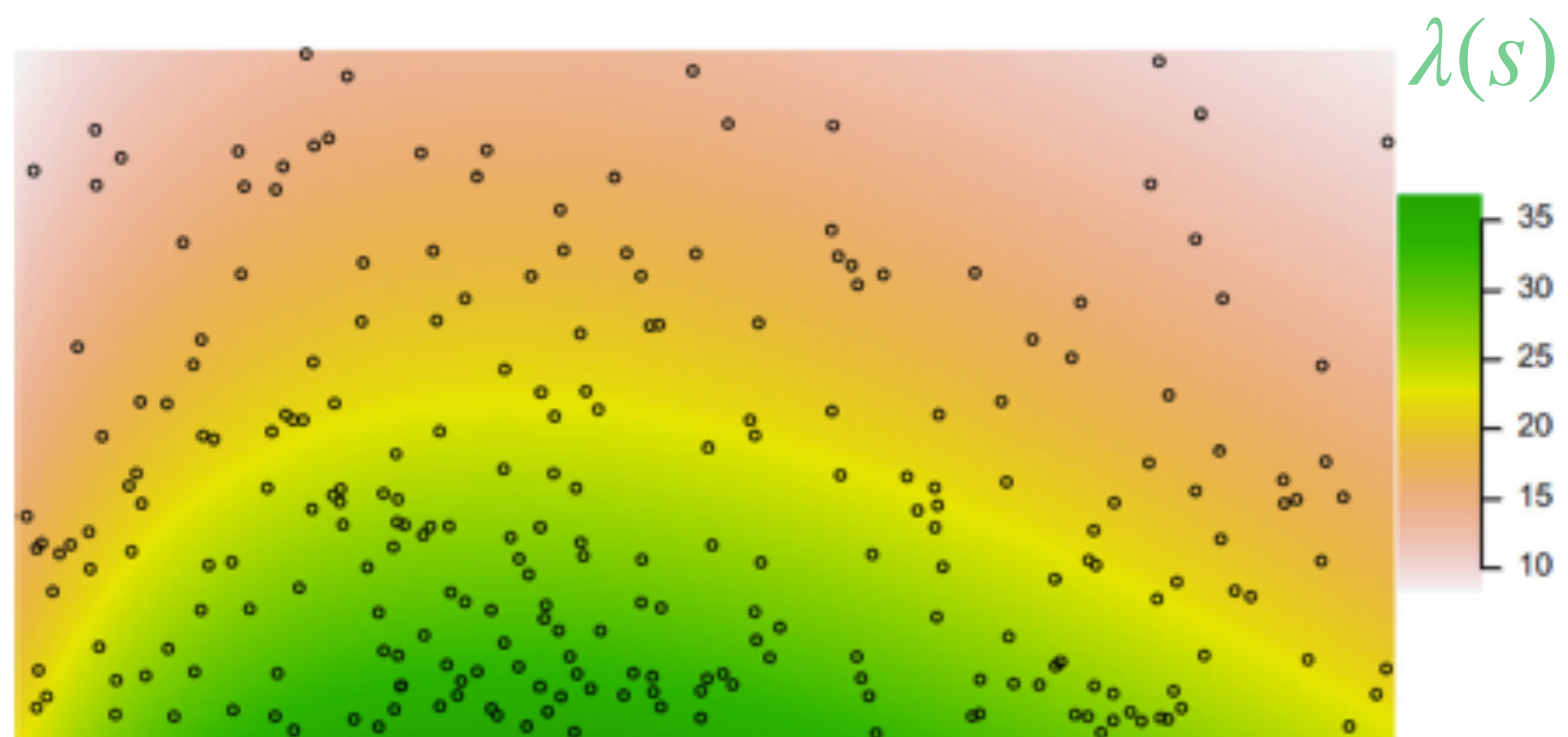
Spotlighting count of possums in 5 min search of 50 m radius.

Restricted spatial extent.

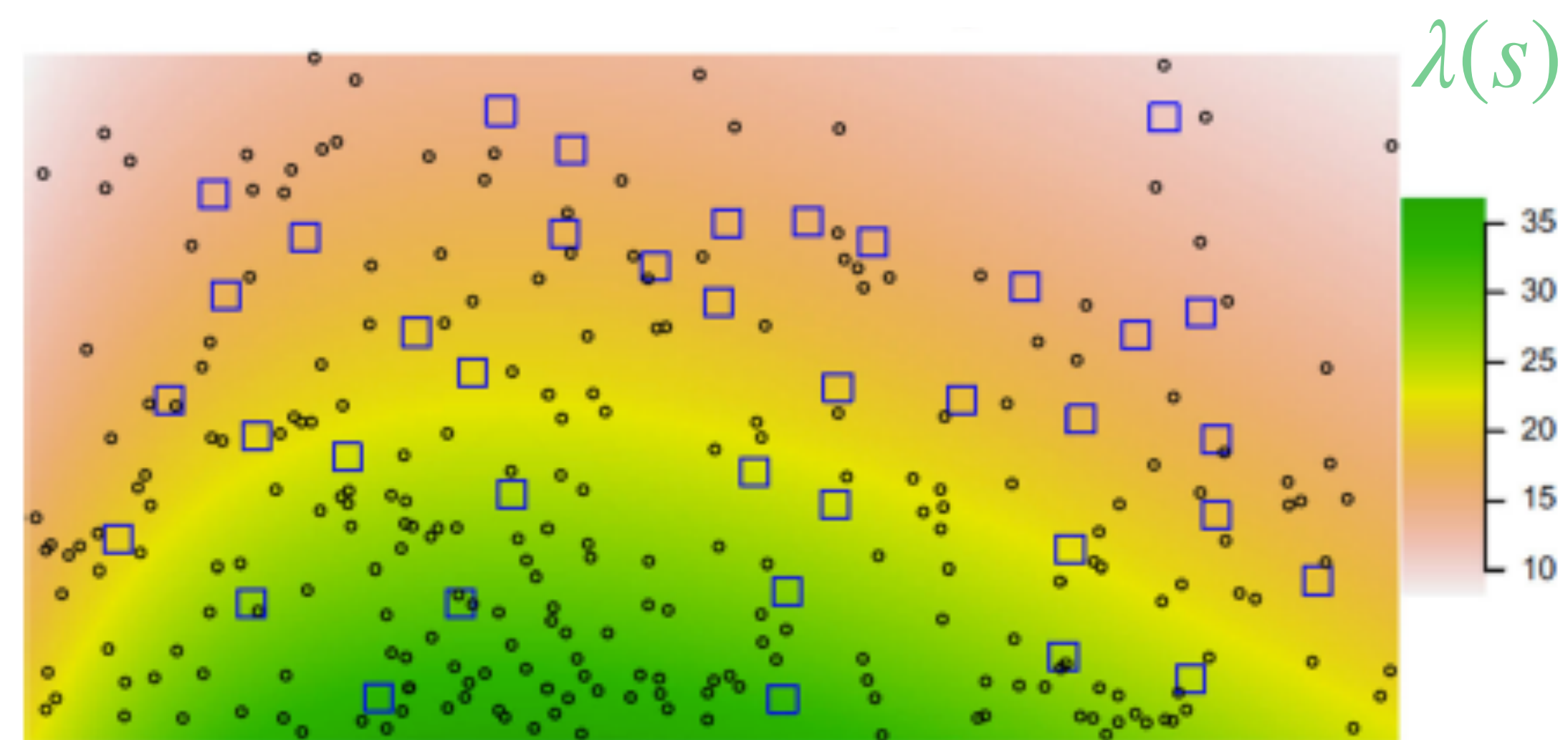
“The truth”

What is the data?





Density across space



**Structured survey
data**

Fithian W, Elith J, Hastie T, Keith DA. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol Evol.* 2015;6(4):424-438. doi:10.1111/2041-210X.12242



count data

What is the model?

$$\textit{count}_i \sim \textit{Poisson}(\lambda_i)$$



count data

What is the model?

$$count_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \alpha + \beta * \text{TreeCover}_i$$

What is the model?

count data

$$count_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \boxed{\alpha} + \boxed{\beta} * \text{TreeCover}_i$$

What is the model?

count data

$$count_i \sim \text{Poisson}(\lambda_i A_{search})$$

$$\log(\lambda_i) = \alpha + \beta * \text{TreeCover}_i$$

50 m search radius = A_{search}



count data

What is the model?

Equivalent to:

$$count_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \alpha + \beta * \text{TreeCover}_i + \log(A_{\text{search}})$$

site
occupancy

What is the data?

Scats detected/not
detected in 10 m radius.

Imprecise proxy for
detection of possums.





site
occupancy

What is the model?

$$occ_k \sim \text{Bernoulli}(\psi_k)$$

$$det_k | occ_k \sim \text{Bernoulli}(occ_k p_k)$$



site
occupancy

What is the model?

$$occ_k \sim \text{Bernoulli}(\psi_k)$$

$$det_k | occ_k \sim \text{Bernoulli}(occ_k p_k)$$

In this case, we are
assuming perfect
detection $p_k = 1$



site
occupancy

What is the model?

$$occ_k \sim \text{Bernoulli}(\psi_k)$$

~~$$det_k | occ_k \sim \text{Bernoulli}(occ_k p_k)$$~~

$$\psi_k = 1 - e^{-Abund_k}$$

so we are modelling ψ_k
instead of p_k

site
occupancy

What is the model?

$$occ_k \sim \textit{Bernoulli}(\psi_k)$$

$$\psi_k = 1 - e^{-Abund_k}$$

“cloglog” link



site
occupancy

What is the model?

$$occ_k \sim \textit{Bernoulli}(\psi_k)$$

$$\psi_k = 1 - e^{-Abund_k}$$

$$Abund_k = \lambda_k * A_{search}$$



site
occupancy

What is the model?

$$occ_k \sim \text{Bernoulli}(\psi_k)$$

$$\psi_k = 1 - e^{-Abund_k}$$

$$Abund_k = \lambda_k * A_{search}$$

$$\lambda_k = \alpha + \beta * TreeCover_k$$



site
occupancy

What is the model?

$$occ_k \sim \text{Bernoulli}(\psi_k)$$

$$\psi_k = 1 - e^{-Abund_k}$$

$$Abund_k = \lambda_k * A_{search}$$

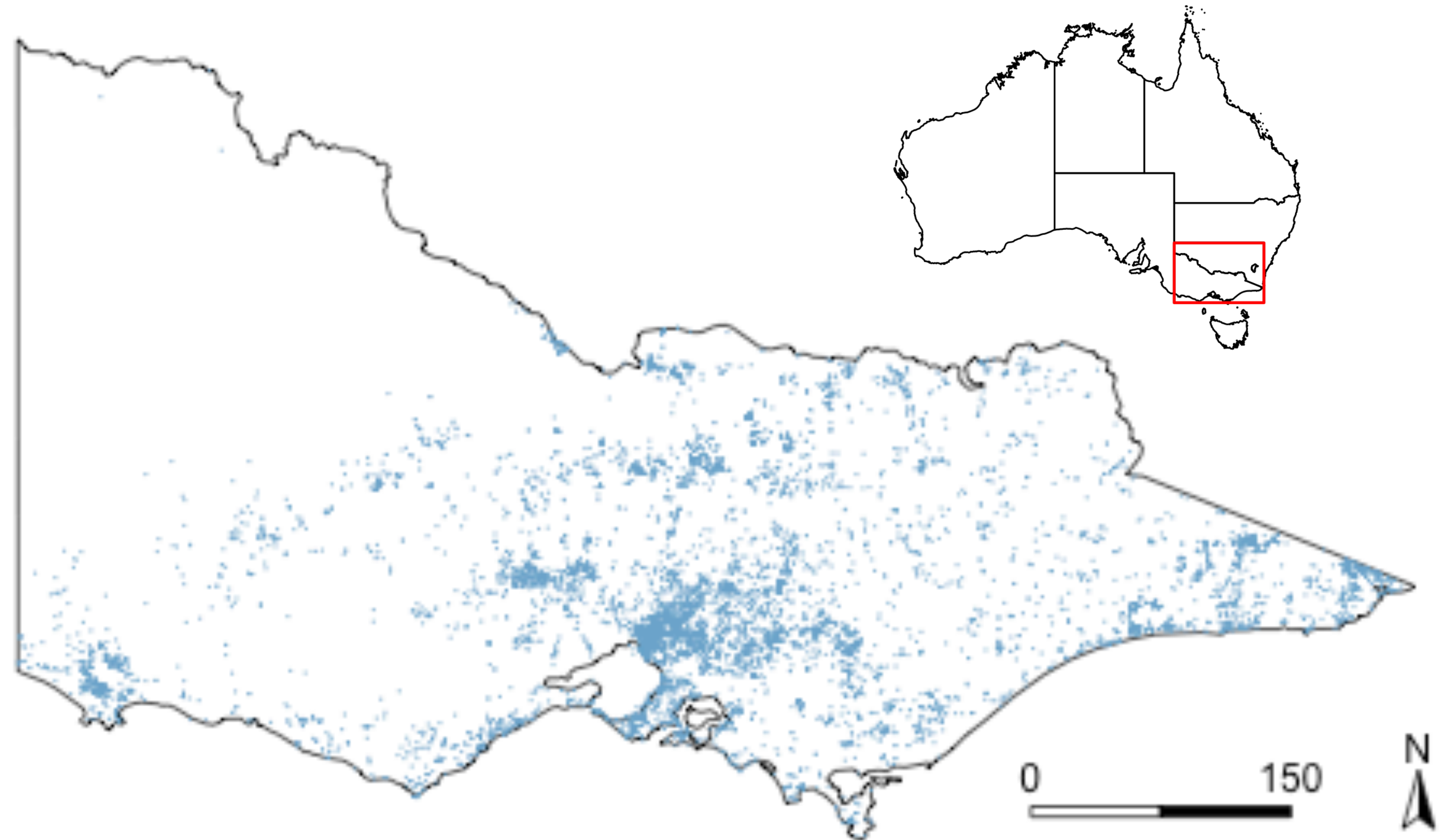
$$\lambda_k = \boxed{\alpha} + \boxed{\beta} * TreeCover_k$$

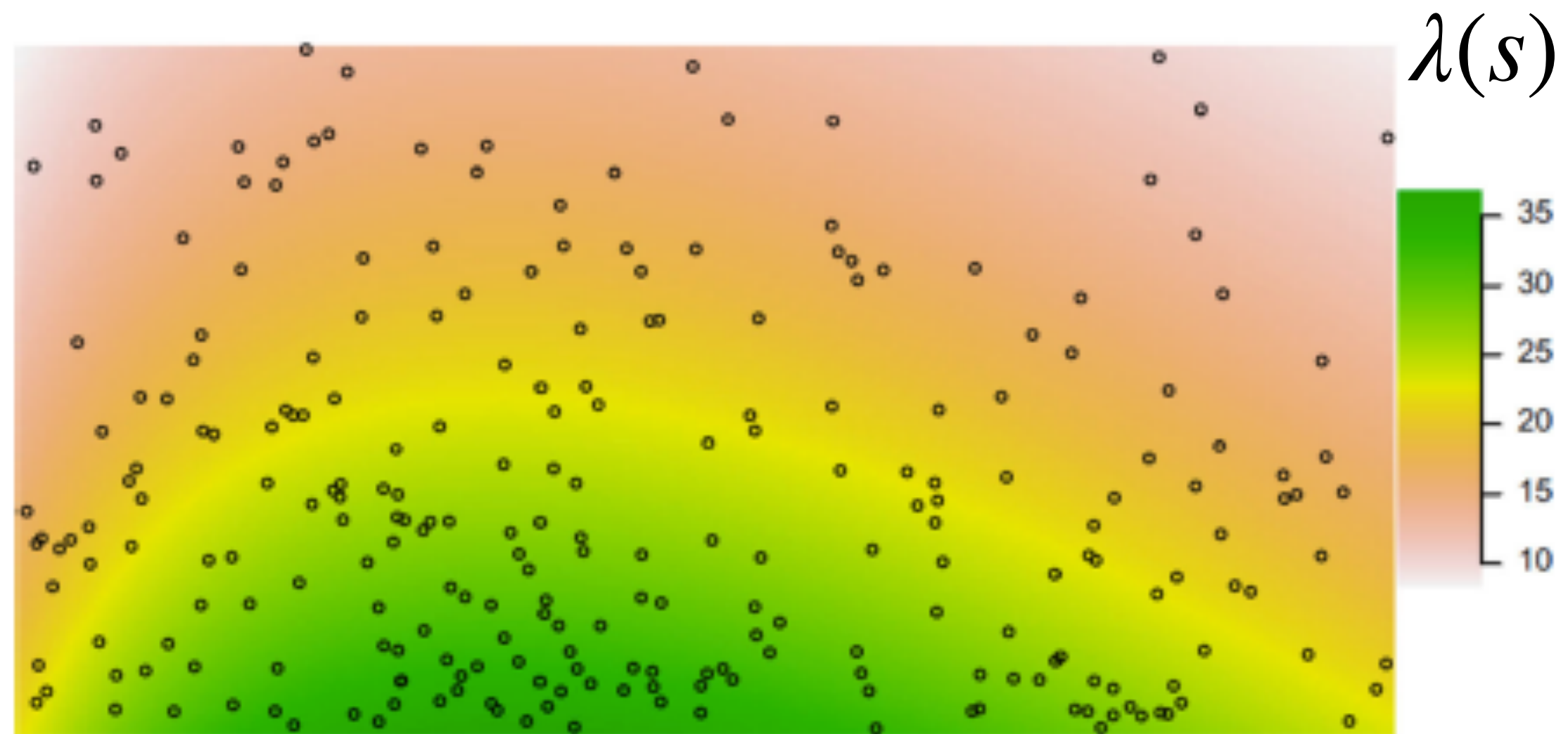
opportunistic observation

Broad scale Atlas of
Living Australia presence
only possum records.

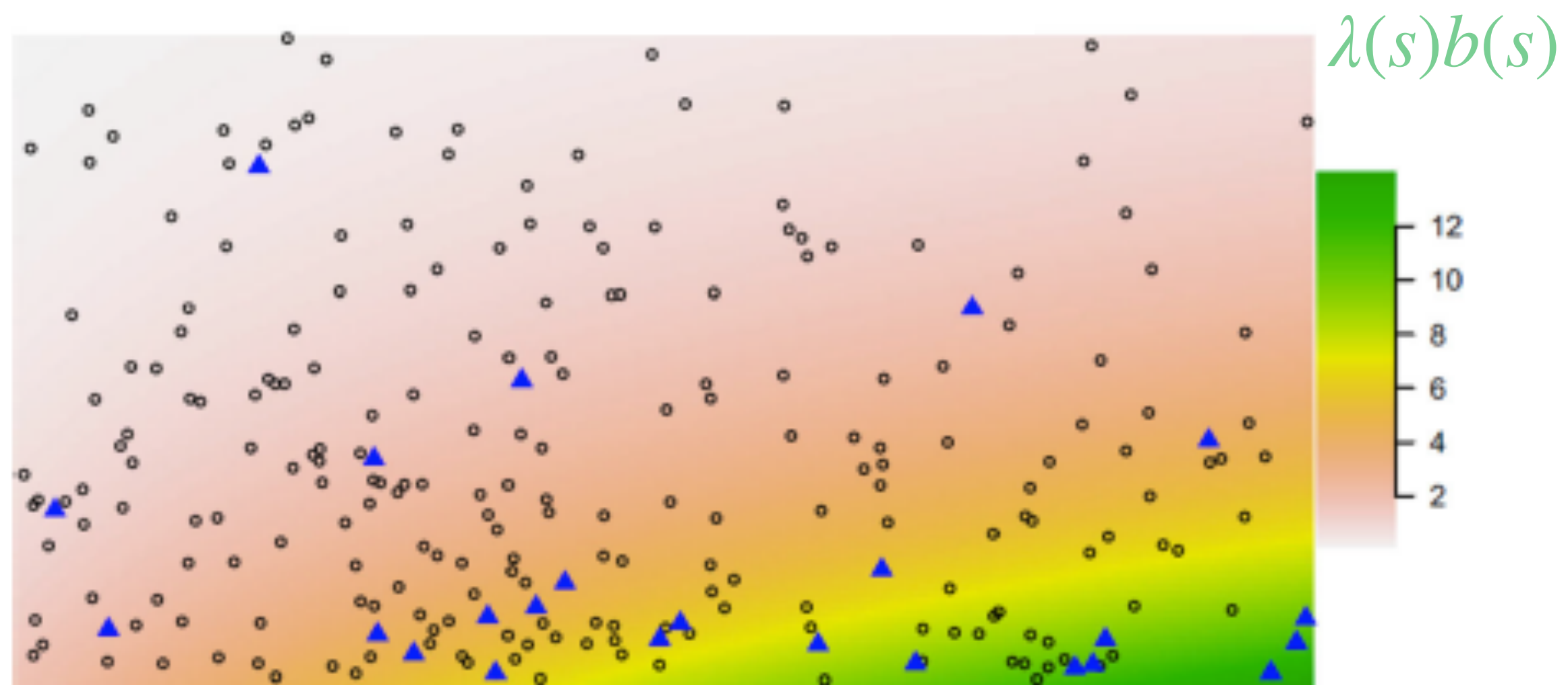
Biased opportunistic/
citizen science data.

What is the data?





Density across space

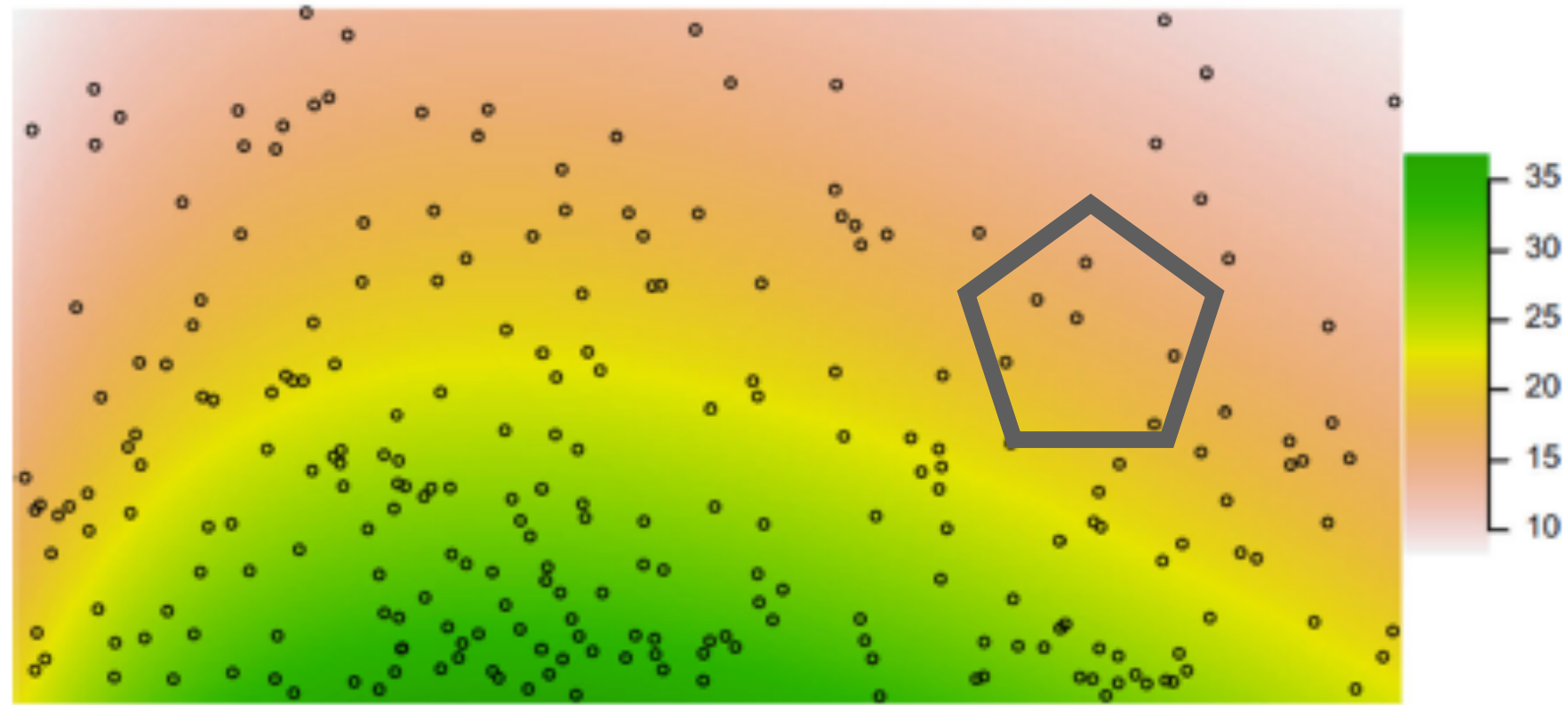


**Biased opportunistic
observations**

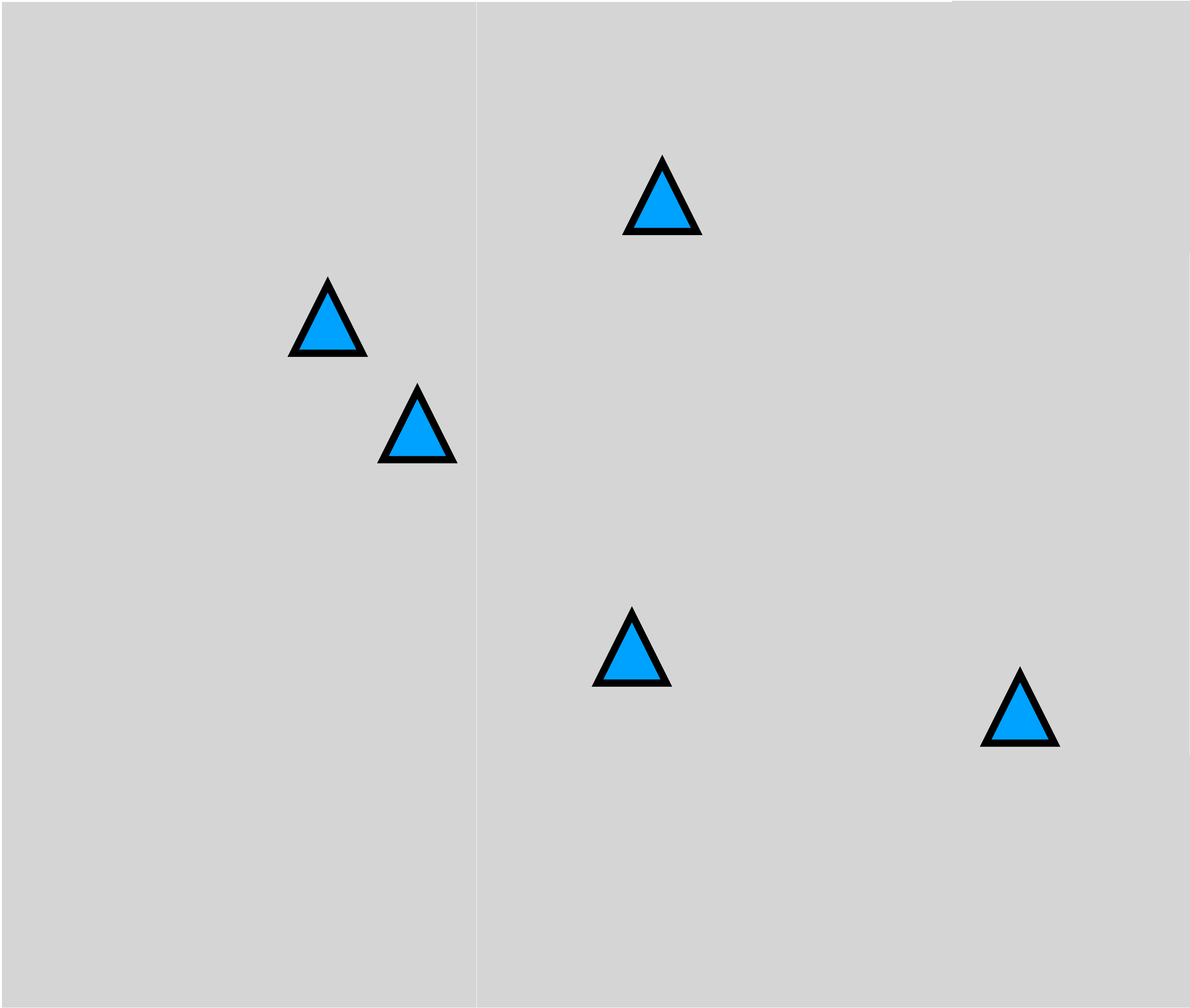
Fithian W, Elith J, Hastie T, Keith DA. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol Evol.* 2015;6(4):424-438. doi:10.1111/2041-210X.12242

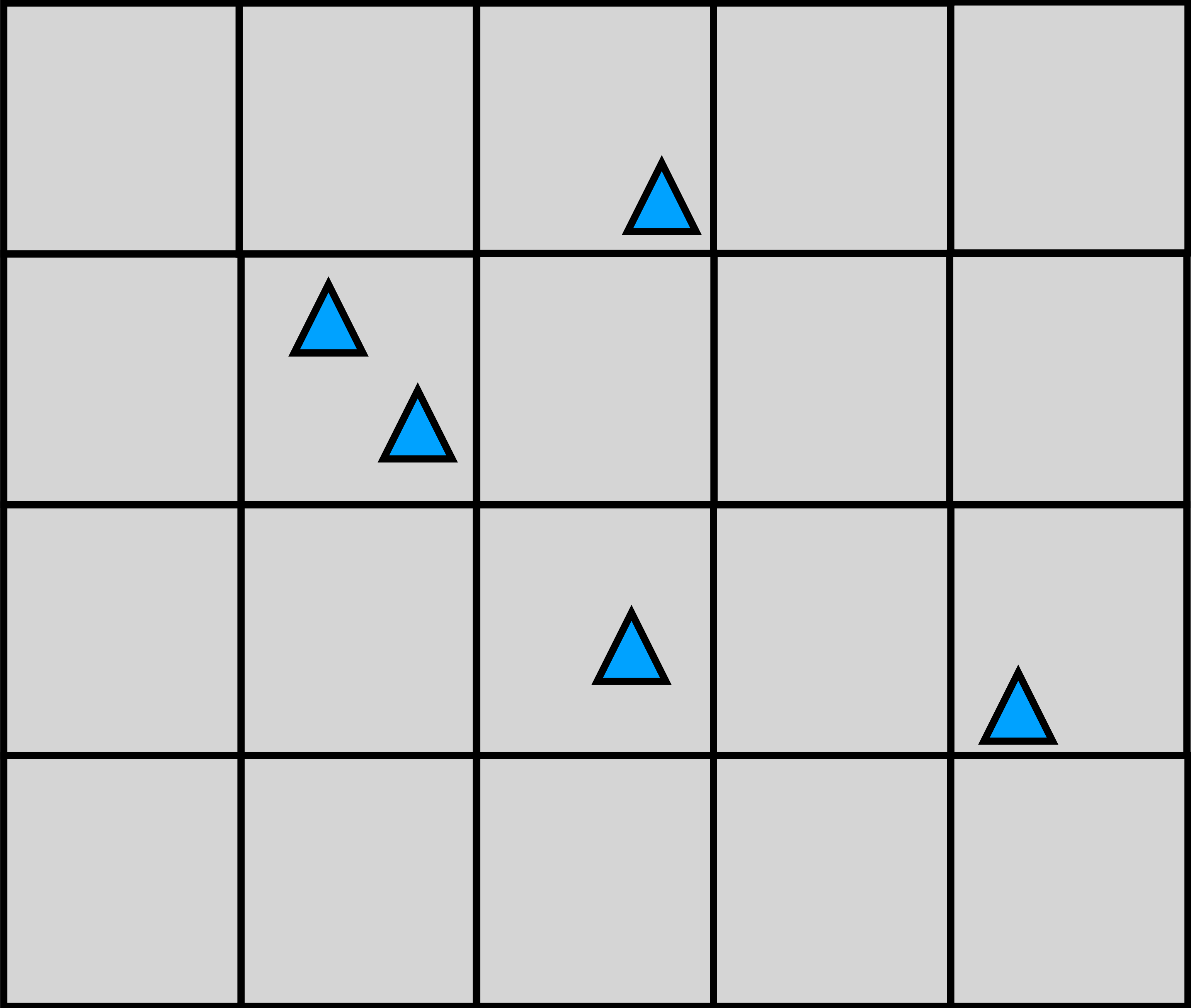
Inhomogenous Poisson point process

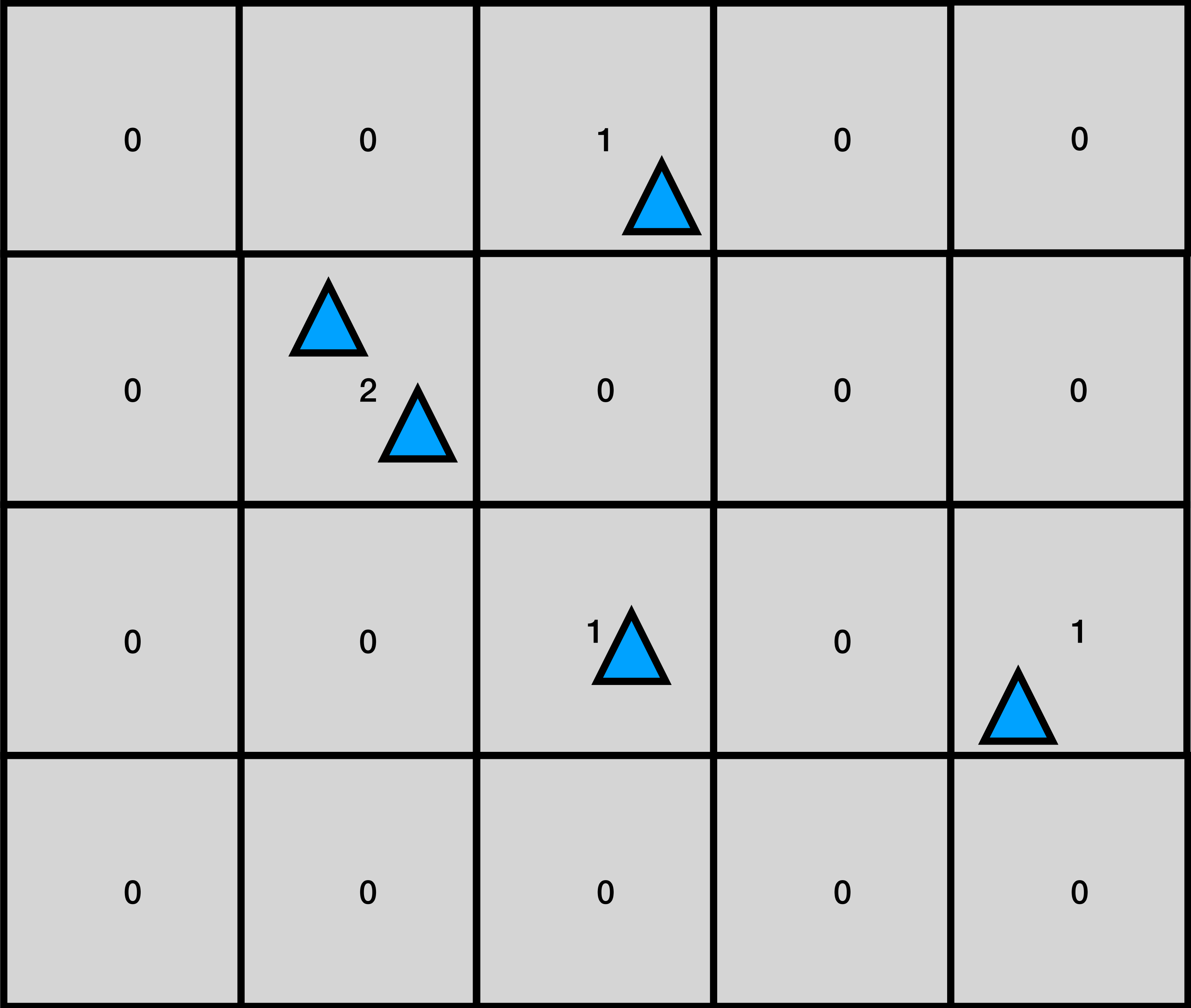
The number of points in any given region is Poisson (regardless of the location, size, or shape of the region)

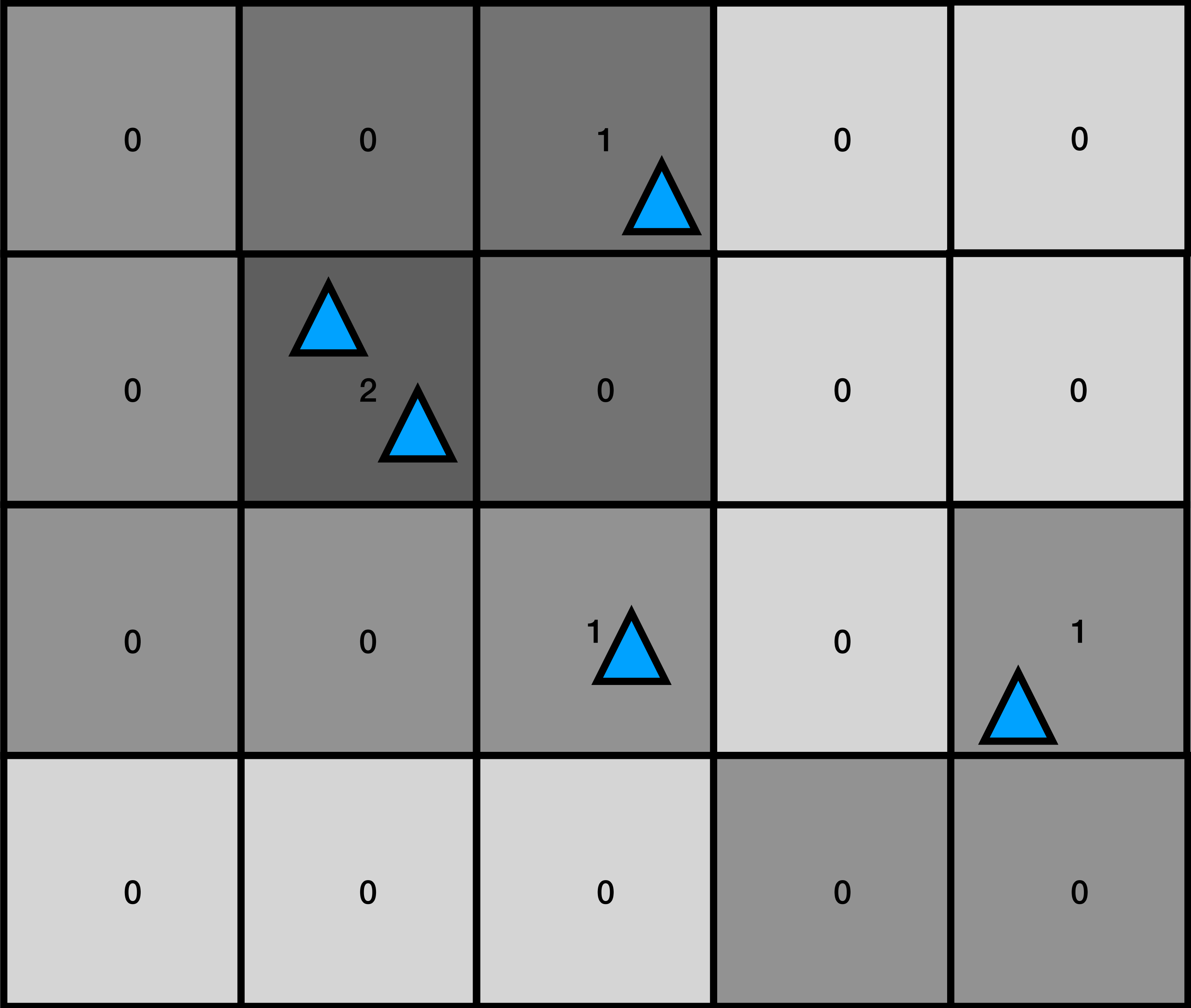


**How to fit IPP
to point data
(cellwise
count
method)**



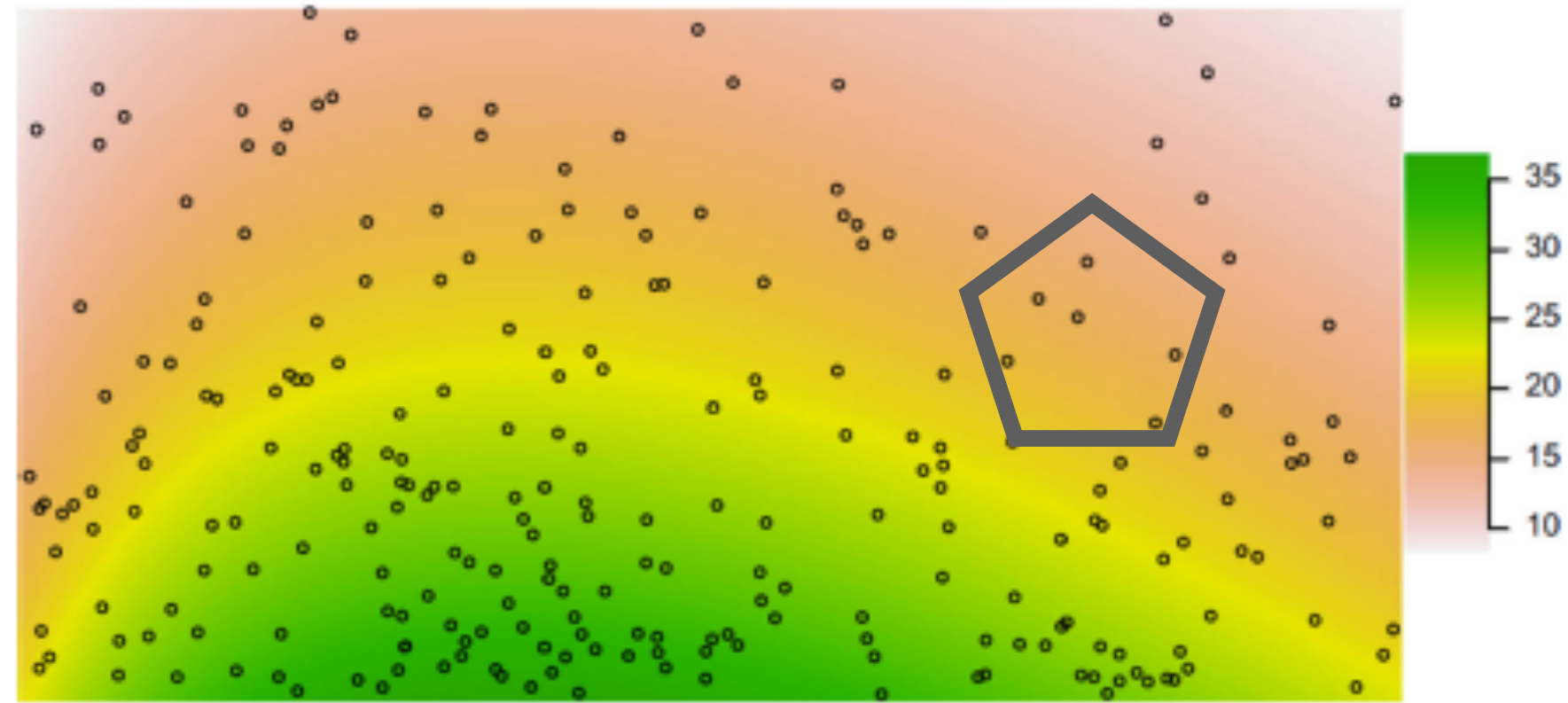






Inhomogenous Poisson point process

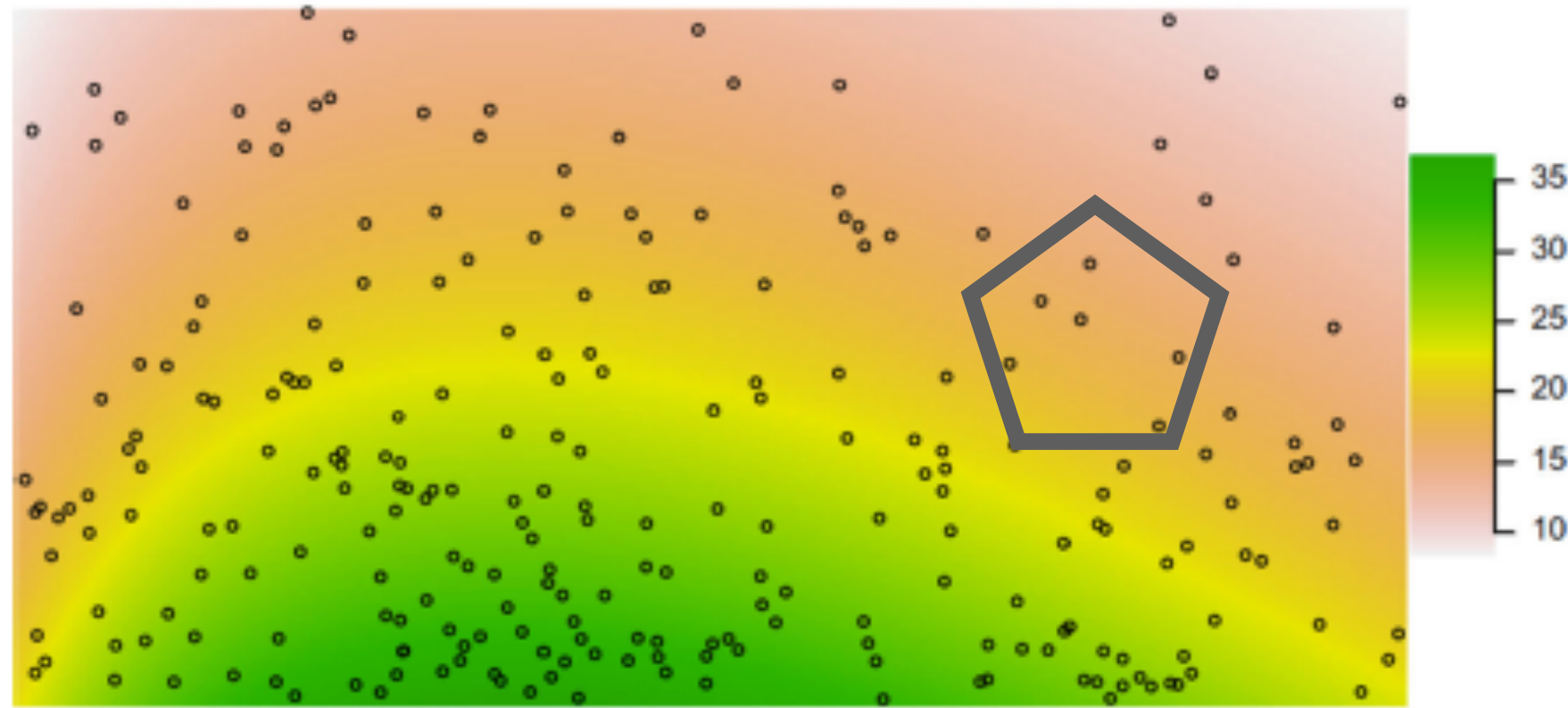
*The number of points in any given region is Poisson
(regardless of the location, size, or shape of the region)*



Expected # points in a region R (parameter of Poisson) =
$$\int_R \lambda(s) ds$$

Inhomogenous Poisson point process

*The number of points in any given region is Poisson
(regardless of the location, size, or shape of the region)*



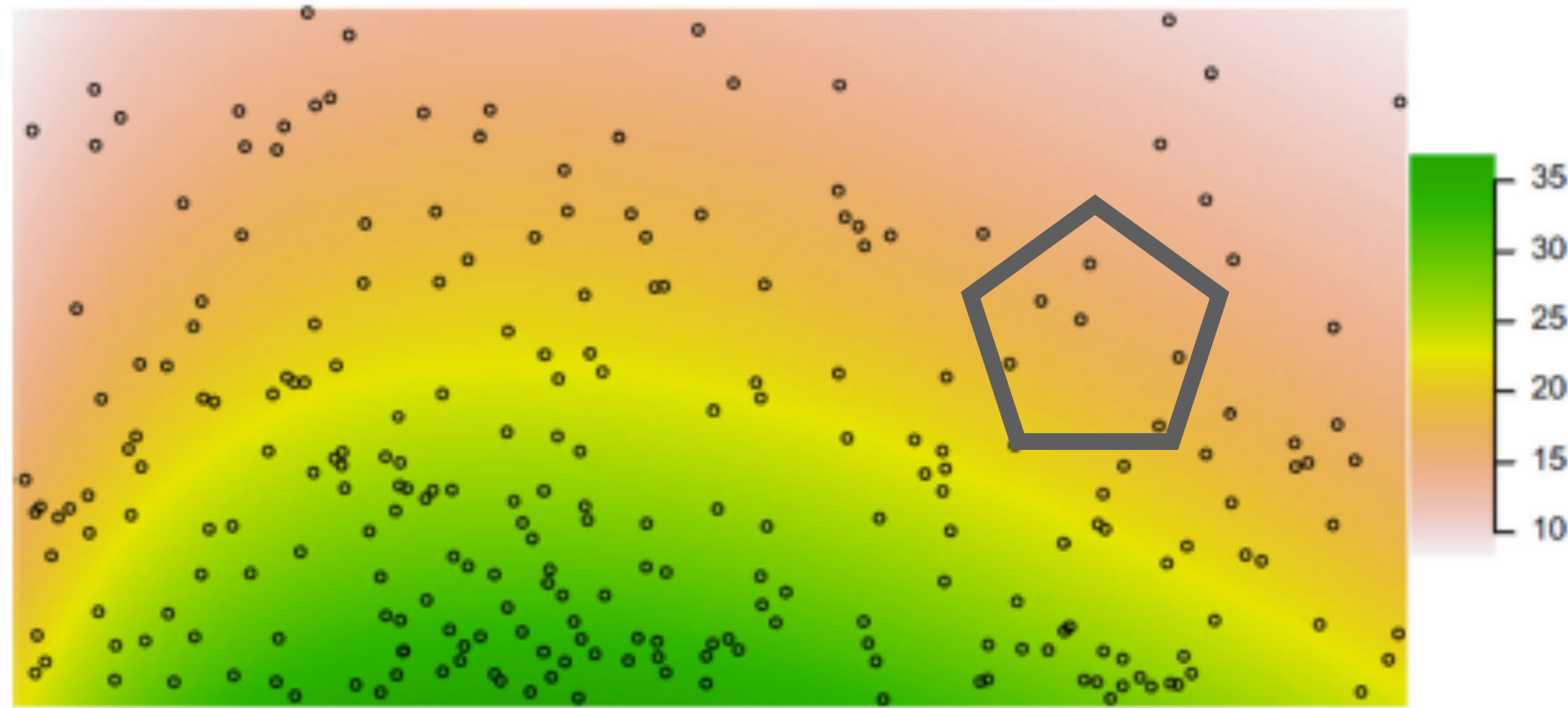
Expected # points in a region R (parameter of Poisson) = $\int_R \lambda(s) ds$

If $\lambda(s)$ is constant over R , with $\lambda(s) = \lambda_R$
and has area A_R then:

$$\frac{\text{Expected number of points in } R}{A_R} =$$

Inhomogenous Poisson point process

*The number of points in any given region is Poisson
(regardless of the location, size, or shape of the region)*



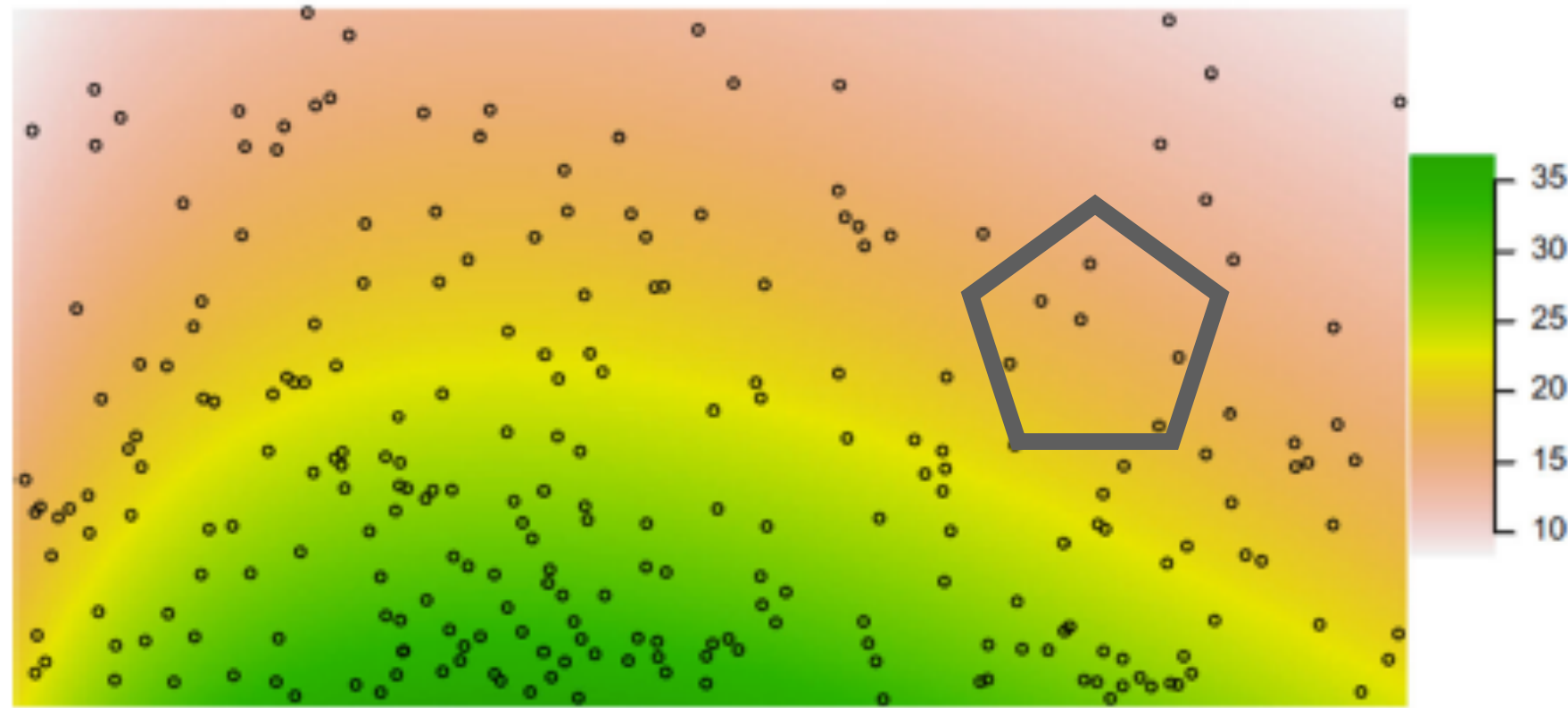
Expected # points in a region R (parameter of Poisson) = $\int_R \lambda(s) ds$

If $\lambda(s)$ is constant over R , with $\lambda(s) = \lambda_R$
and has area A_R then:

$$\frac{\text{Expected number of points in } R}{A_R} = \lambda_R = \text{density of points (same units as } A_R)$$

Inhomogenous Poisson point process

The number of points in any given region is Poisson (regardless of the location, size, or shape of the region)



Expected # points in a region R (parameter of Poisson) = $\int_R \lambda(s) ds$

If $\lambda(s)$ is constant over R , with $\lambda(s) = \lambda_R$ and has area A_R then:

$$\text{Expected \# points in } R = \lambda_R A_R$$

$$\frac{\text{Expected number of points in } R}{A_R} = \lambda_R = \text{density of points (same units as } A_R)$$

The possum density process model linking all data types

$$\begin{array}{l} \text{Expected \# possums} \\ \text{in region } i \end{array} = \int_i \lambda(s) ds \approx \lambda_i a_i$$

The possum density process model linking all data types

$$\begin{array}{l} \text{Expected \# possums} \\ \text{in region } i \end{array} = \int_i \lambda(s) ds \approx \lambda_i a_i$$

$$\text{Possum density in } i : \log(\lambda_i) = \alpha + \beta * \textit{TreeCover}_i$$

The possum density process model linking all data types

Expected # possums
in region i

$$= \int_i \lambda(s) ds \approx \lambda_i a_i$$

Possum density in i : $\log(\lambda_i) = \boxed{\alpha} + \boxed{\beta} * TreeCover_i$

**opportunistic
observation**

What is the model?

$$po \sim IPP(\lambda(s)b(s))$$

**opportunistic
observation**

What is the model?

$$po \sim IPP(\lambda(s)b(s))$$

$$po_j \sim Poisson(\lambda_j b_j A_j)$$

opportunistic
observation

What is the model?

$$po \sim IPP(\lambda(s)b(s))$$

$$po_j \sim Poisson(\lambda_j b_j A_j)$$

$$\log(\lambda_j) = \alpha + \beta * TreeCover_j$$

opportunistic
observation

What is the model?

$$po \sim IPP(\lambda(s)b(s))$$

$$po_j \sim Poisson(\lambda_j b_j A_j)$$

$$\log(\lambda_j) = \boxed{\alpha} + \boxed{\beta} * TreeCover_j$$

**opportunistic
observation**

What is the model?

$$po \sim IPP(\lambda(s)b(s))$$

$$po_j \sim Poisson(\lambda_j b_j A_j)$$

$$\log(\lambda_j) = \alpha + \beta * TreeCover_j$$

$$\log(b_j) = \alpha_{bias} + \beta_{bias} CityAccess_j$$

**opportunistic
observation**

What is the model?

$$po \sim IPP(\lambda(s)b(s))$$

Equivalent to:

$$po_j \sim Poisson(\Lambda_j A_j)$$

$$\log(\Lambda_j) = \alpha + \beta * TreeCover_j + \alpha_{bias} + \beta_{bias} CityAccess_j$$

We do not need to fit these in JAGS, can be fit in glm framework, but we want to make the maths explicit.

Over to the code....

Thank you!

More reading:

Fithian W, Elith J, Hastie T, Keith DA. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol Evol*. 2015;6(4):424-438. doi:10.1111/2041-210X.12242

Guillera-Arroita, G. (2017), Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, 40: 281-295. <https://doi.org/10.1111/ecog.02445>

Nick J.B. Isaac, Marta A. Jarzyna, Petr Keil, Lea I. Dambly, Philipp H. Boersch-Supan, Ella Browning, Stephen N. Freeman, Nick Golding, Gurutzeta Guillera-Arroita, Peter A. Henrys, Susan Jarvis, José Lahoz-Monfort, Jörn Pagel, Oliver L. Pescott, Reto Schmucki, Emily G. Simmonds, Robert B. O'Hara. 2020.

Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology & Evolution*, 35:1, 56-67, <https://doi.org/10.1016/j.tree.2019.08.006>.

Simmonds, E.G., Jarvis, S.G., Henrys, P.A., Isaac, N.J.B. and O'Hara, R.B. (2020), Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, 43: 1413-1422. <https://doi.org/10.1111/ecog.05146>



Australian Government
National Health and Medical Research Council

