The rise and rise of joint species distribution models (JSDMs) in ecology

Francis K.C. Hui Australian National University

Multivariate abundance data

Gen 1: MGLMM

Gen 2: Latent variables/factor analysis

Gen 2+: LVMs with all the extras

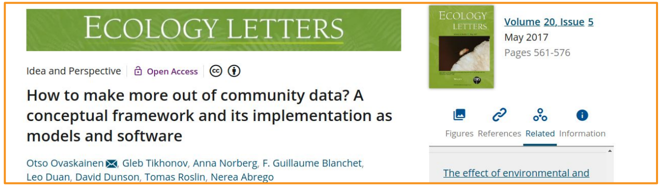
Closing thoughts

Some images courtesy of SlidesCarnival and Google images
AASC 2022

Disclaimer

- This is an opinionated review/perspective talk, so you will see a decent chunk of my and my collaborators' works
  - Apologies for this!
  - Thank you to all who have/continue to inspire me to work on JSDMs

Multivariate abundance data



ECOLOGY LETTERS

Idea and Perspective | Open Access

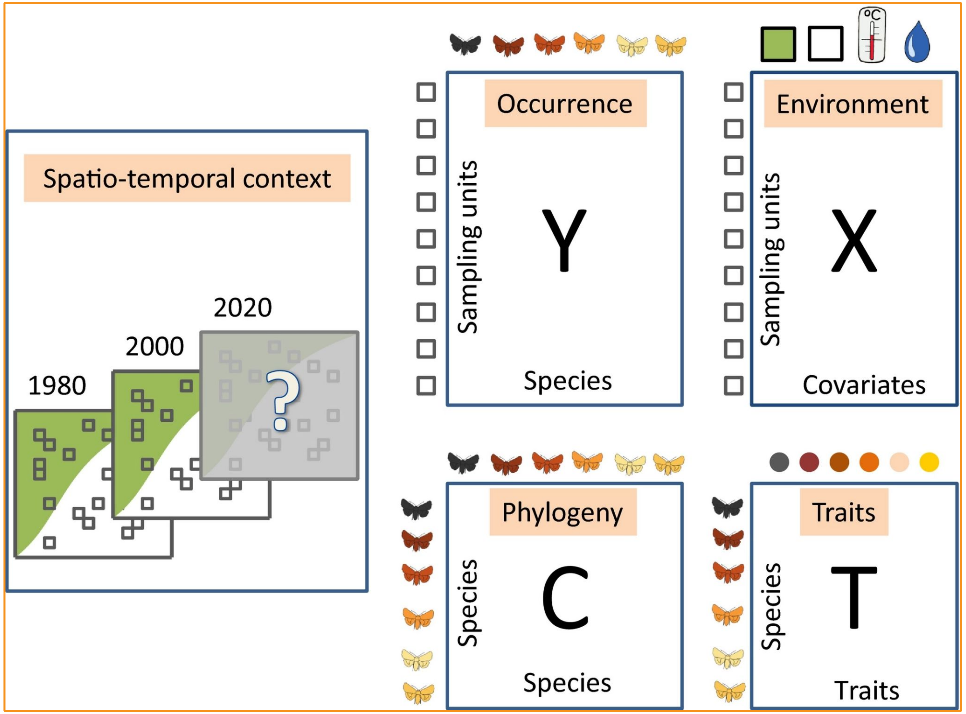
How to make more out of community data? A conceptual framework and its implementation as models and software

Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, Nerea Abrego

Volume 20, Issue 5
May 2017
Pages 561-576

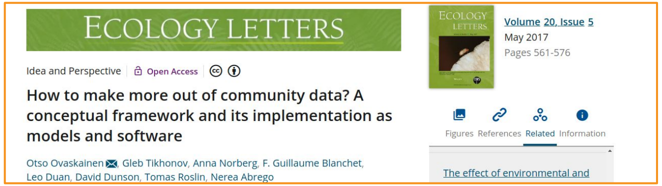
Figures References Related Information

The effect of environmental and



Multivariate abundance data

- Some common features:
  - ○ Multiple correlated responses (high-dimensional)
  - ○ Non-continuous responses with evident mean-variance relationship
  - ○ Non-linear Y-X relationships



ECOLOGY LETTERS

Idea and Perspective | Open Access

How to make more out of community data? A conceptual framework and its implementation as models and software

Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, Nerea Abrego

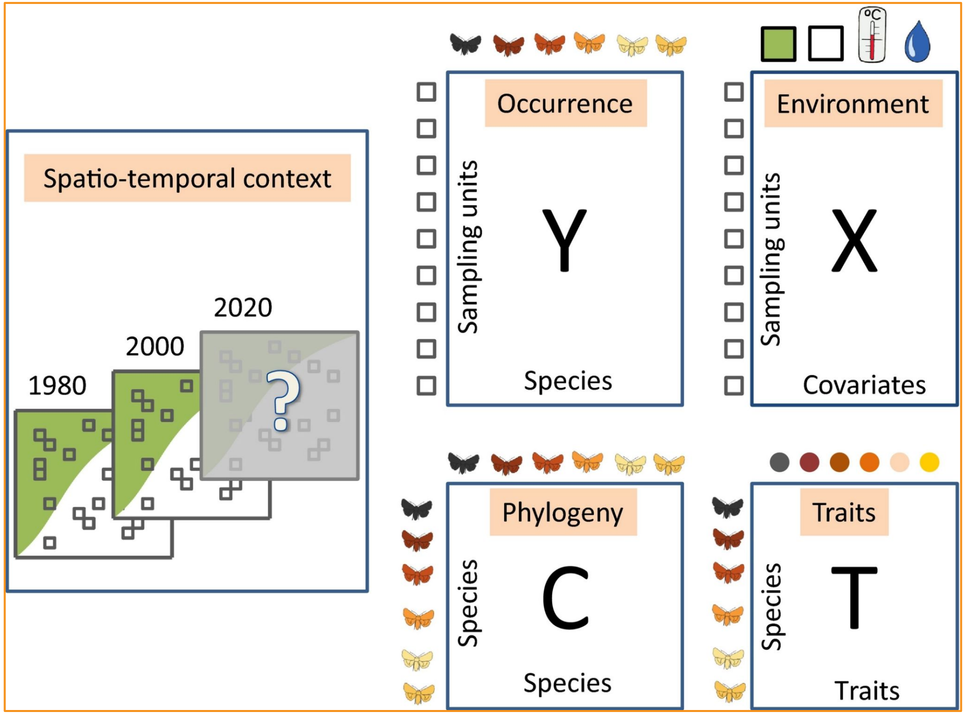
Volume 20, Issue 5

May 2017

Pages 561-576

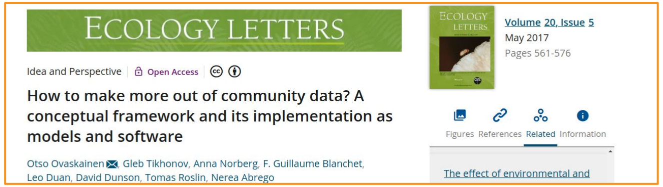
Figures References Related Information

The effect of environmental and



Question/s of interests

- Depends on the data you have:
  - (a) is a multivariate prediction problem
  - (b) -> how is Y and X related?
  - (c) -> how are the columns of Y related?
  - (d) + (e) -> how do T & C mediate/drive the Y-X relationship?



ECOLOGY LETTERS

Idea and Perspective | Open Access

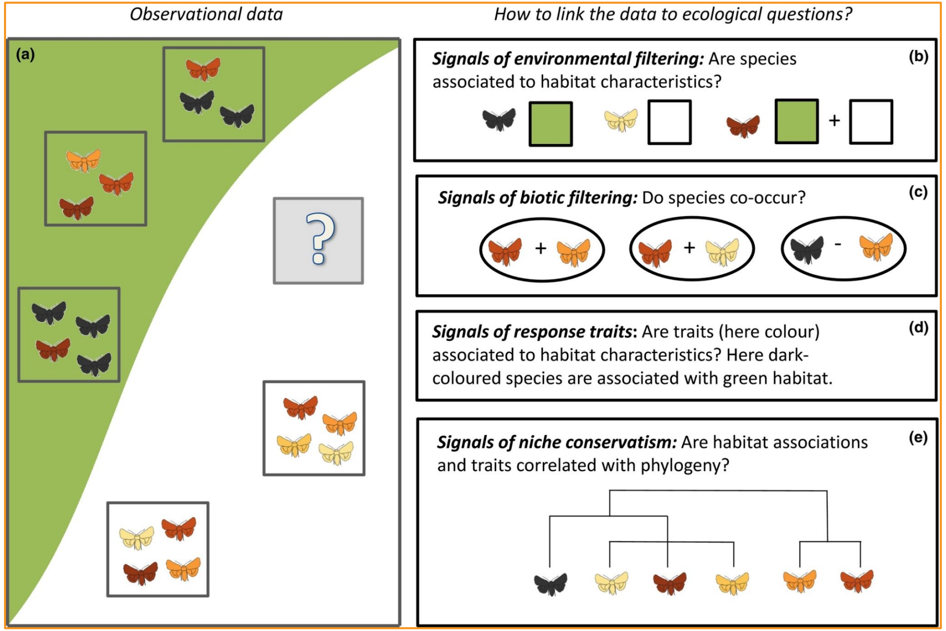
How to make more out of community data? A conceptual framework and its implementation as models and software

Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, Nerea Abrego

Volume 20, Issue 5

May 2017

Pages 561-576



Observational data

How to link the data to ecological questions?

Signals of environmental filtering: Are species associated to habitat characteristics?

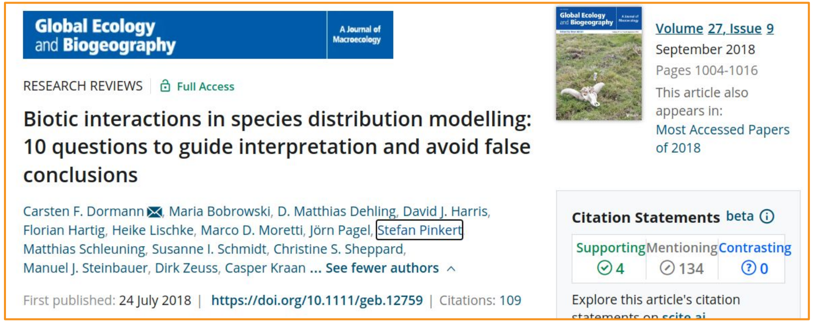
Signals of biotic filtering: Do species co-occur?

Signals of response traits: Are traits (here colour) associated to habitat characteristics? Here dark-coloured species are associated with green habitat.

Signals of niche conservatism: Are habitat associations and traits correlated with phylogeny?

Enter the joint species distribution model

- Loosely speaking, a joint species distribution model (JSDM) refers to a statistical method that simultaneously models all species
    - Accounts for the fact that species may be correlated with each other (after adjusting for measured predictor)
    - A single, potentially high-dimensional log-likelihood function
    - The sources of this (residual) correlation could be many…



Global Ecology and Biogeography | A Journal of Macroecology

RESEARCH REVIEWS | Full Access

Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions

Carsten F. Dormann, Maria Bobrowski, D. Matthias Dehling, David J. Harris, Florian Hartig, Heike Lischke, Marco D. Moretti, Jörn Pagel, Stefan Pinkert Matthias Schleuning, Susanne I. Schmidt, Christine S. Sheppard, Manuel J. Steinbauer, Dirk Zeuss, Casper Kraan ... See fewer authors

First published: 24 July 2018 | https://doi.org/10.1111/geb.12759 | Citations: 109

Volume 27, Issue 9

September 2018

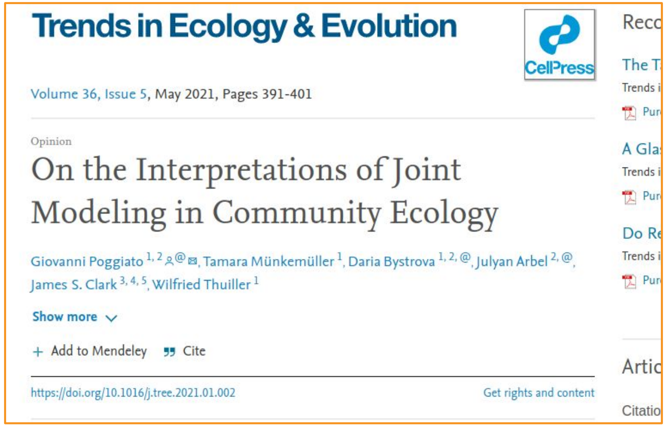
Pages 1004-1016

This article also appears in: Most Accessed Papers of 2018

Citation Statements beta

Supporting	Mentioning	Contrasting
4	134	0

Explore this article's citation statements on scite.ai



Trends in Ecology & Evolution

Volume 36, Issue 5, May 2021, Pages 391-401

Opinion

On the Interpretations of Joint Modeling in Community Ecology

Giovanni Poggiato[1,2], Tamara Münkemüller[1], Daria Bystrova[1,2], Julyan Arbel[2], James S. Clark[3,4,5], Wilfried Thuiller[1]

Show more

https://doi.org/10.1016/j.tree.2021.01.002 Get rights and content

Enter the joint species distribution model

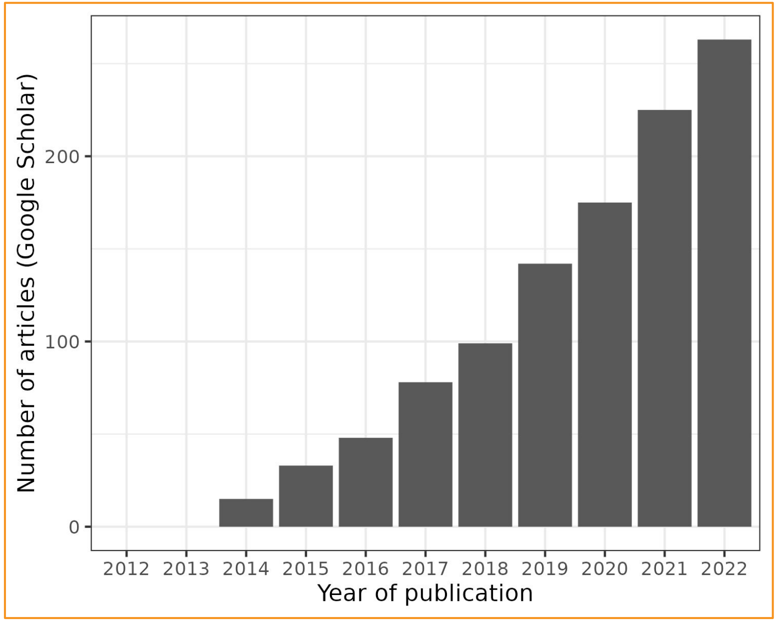
- Loosely speaking, a joint species distribution model (JSDM) refers to a statistical method that simultaneously models all species
    - Accounts for the fact that species may be correlated with each other (after adjusting for measured predictor)
    - A single, potentially high-dimensional log-likelihood function
    - The sources of this (residual) correlation could be many…

- JSDMs are basically a counterpart to stacked species distribution models (SSDMs), which model each species separately
    - Log-likelihood function comprises the sum of independent species contributions e.g., fit a GLM/GAM/GLMM/ML etc…to each species

Enter the joint species distribution model

- A Google Scholar search of four key JSDM phrases (as of 23 November 2022)
  - `Joint species distribution models`
  - `Model-based ordination`
  - `Joint dynamic species distribution models`
  - `Hierarchical modeling of species communities`

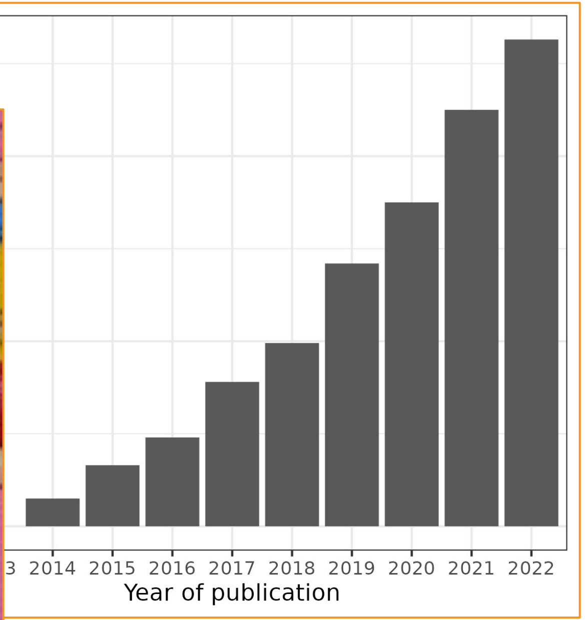
- This is probably an underestimate of JSDM's rise...



Enter the joint species distribution model

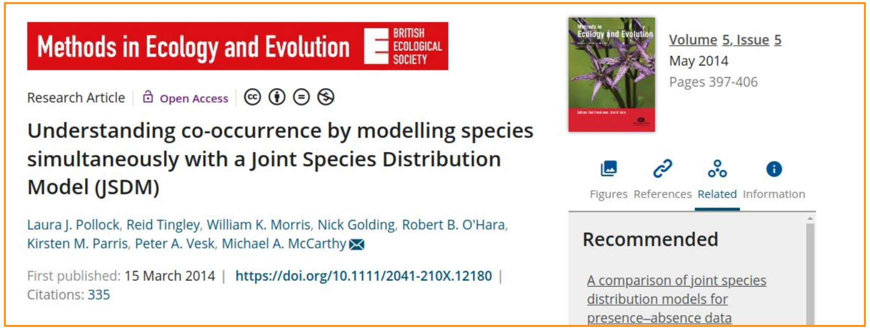
- A Google Scholar search of four key JSDM phrases (as of 8 October 2022)
    - `Joint spec`
    - `Model-base`
    - `Joint dyna models`
    - `Hierarchic communitie`

- This is probably JSDM's rise…



Gen 1: MGLMMs

- Multivariate generalized linear mixed model (MGLMM)
  - Model residual between-species correlations using a multivariate random intercept



Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Volume 5, Issue 5
May 2014
Pages 397-406

Research Article | Open Access

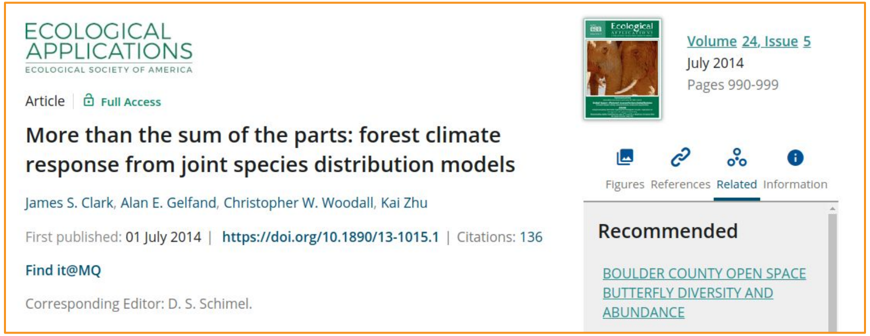
Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)

Laura J. Pollock, Reid Tingley, William K. Morris, Nick Golding, Robert B. O'Hara, Kirsten M. Parris, Peter A. Vesk, Michael A. McCarthy

First published: 15 March 2014 | https://doi.org/10.1111/2041-210X.12180 | Citations: 335

Recommended

A comparison of joint species distribution models for presence–absence data



ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA

Volume 24, Issue 5
July 2014
Pages 990-999

Article | Full Access

More than the sum of the parts: forest climate response from joint species distribution models

James S. Clark, Alan E. Gelfand, Christopher W. Woodall, Kai Zhu

First published: 01 July 2014 | https://doi.org/10.1890/13-1015.1 | Citations: 136

Find it@MQ

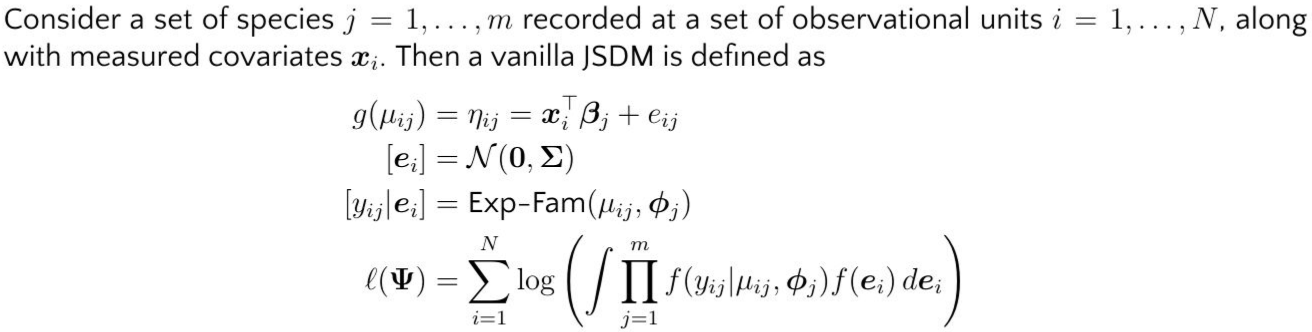
Corresponding Editor: D. S. Schimel.

Recommended

BOULDER COUNTY OPEN SPACE BUTTERFLY DIVERSITY AND ABUNDANCE

Gen 1: MGLMMs

- Multivariate generalized linear mixed model (MGLMM)
  - Model residual between-species correlations using a multivariate random intercept
  - Exponential family is being used “loosely” here to cover many response distributions

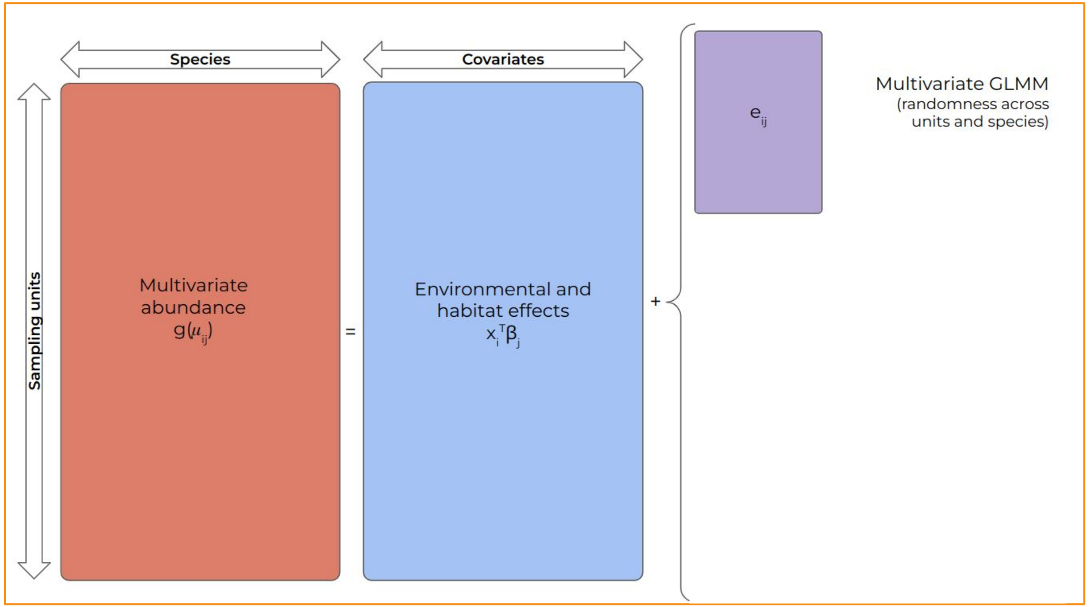


Consider a set of species $j = 1, \ldots, m$ recorded at a set of observational units $i = 1, \ldots, N$, along with measured covariates $\boldsymbol{x}_i$. Then a vanilla JSDM is defined as

$$\begin{aligned} g(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + e_{ij} \\ [\boldsymbol{e}_i] &= \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ [y_{ij}|\boldsymbol{e}_i] &= \text{Exp-Fam}(\mu_{ij}, \boldsymbol{\phi}_j) \\ \ell(\boldsymbol{\Psi}) &= \sum_{i=1}^{N} \log \left(\int \prod_{j=1}^{m} f(y_{ij}|\mu_{ij}, \boldsymbol{\phi}_j) f(\boldsymbol{e}_i) \, d\boldsymbol{e}_i \right) \end{aligned}$$

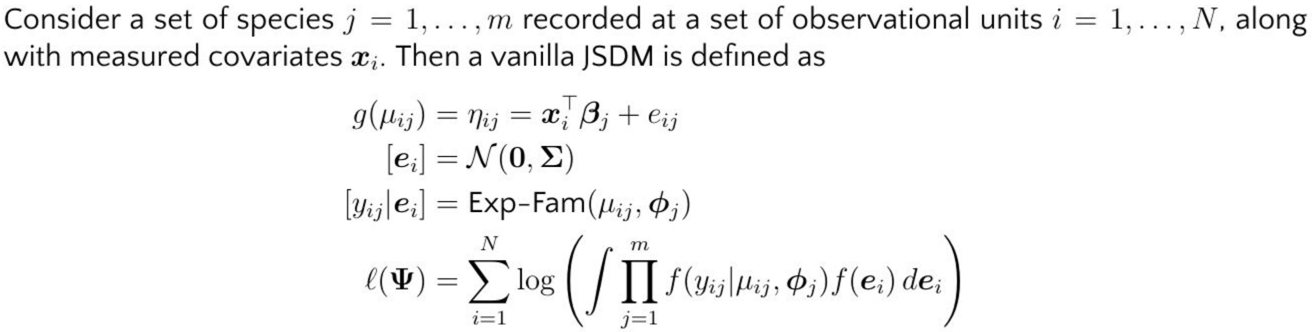
Gen 1: MGLMMs

- Multivariate generalized linear mixed model (MGLMM)
  - Model residual between-species correlations using a multivariate random intercept
  - Exponential family is being used “loosely” here to cover many response distributions


Species
Covariates
Sampling units
Multivariate abundance $g(\mu_{ij})$
=
Environmental and habitat effects $x_i^T\beta_j$
+
e_{ij}
Multivariate GLMM
(randomness across units and species)

Gen 1: MGLMMs

- Multivariate generalized linear mixed model (MGLMM)
  - Model residual between-species correlations using a multivariate random intercept
  - Exponential family is being used “loosely” here to cover many response distributions
  - Pretty flexible (at least for correlations/symmetric associations)
  - Number of parameters scale as m^2, so great if m is not large (compared to N)
  - Lots of random effects, scaling as Nm



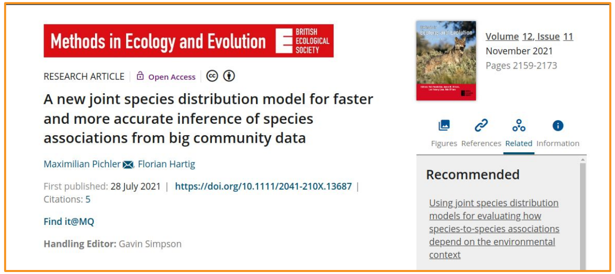
Consider a set of species $j = 1, \ldots, m$ recorded at a set of observational units $i = 1, \ldots, N$, along with measured covariates $\boldsymbol{x}_i$. Then a vanilla JSDM is defined as

$$
\begin{aligned}
g(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + e_{ij} \\
[\boldsymbol{e}_i] &= \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\
[y_{ij} | \boldsymbol{e}_i] &= \text{Exp-Fam}(\mu_{ij}, \boldsymbol{\phi}_j) \\
\ell(\boldsymbol{\Psi}) &= \sum_{i=1}^{N} \log \left(\int \prod_{j=1}^{m} f(y_{ij} | \mu_{ij}, \boldsymbol{\phi}_j) f(\boldsymbol{e}_i) \, d\boldsymbol{e}_i \right)
\end{aligned}
$$

Gen 1: MGLMMs

- Multivariate generalized linear mixed model (MGLMM)
    - Model residual between-species correlations using a multivariate random intercept
    - Exponential family is being used “loosely” here to cover many response distributions
    - Pretty flexible (at least for correlations/symmetric associations)
    - Number of parameters scale as m^2, so great if m is not large (compared to N)
    - Lots of random effects, scaling as Nm

- Largely overtaken by Gen 2 JSDMs, but advances continue to be made…
    - Translating ideas from sparse graphical model/network/ML literature



RESEARCH ARTICLE | Open Access

A new joint species distribution model for faster and more accurate inference of species associations from big community data

Maximilian Pichler, Florian Hartig

First published: 28 July 2021 | https://doi.org/10.1111/2041-210X.13687 | Citations: 5

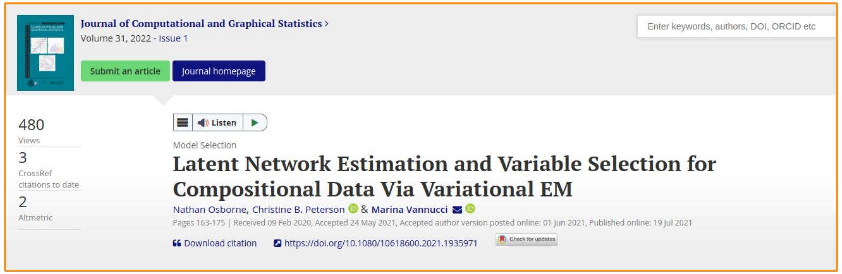
Find it@MQ

Handling Editor: Gavin Simpson

Volume 12, Issue 11
November 2021
Pages 2159-2173

Recommended

Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context



Journal of Computational and Graphical Statistics ›
Volume 31, 2022 - Issue 1

480
Views
3
CrossRef citations to date
2
Altmetric

Model Selection

Latent Network Estimation and Variable Selection for Compositional Data Via Variational EM

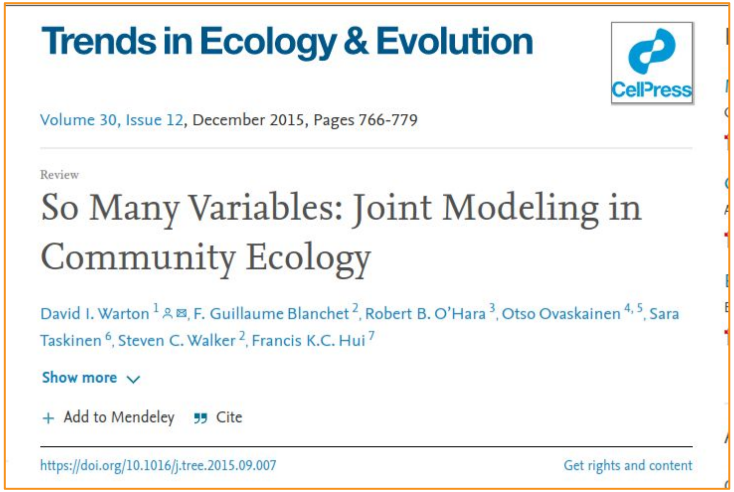
Nathan Osborne, Christine B. Peterson & Marina Vannucci

Pages 163-175 | Received 09 Feb 2020, Accepted 24 May 2021, Accepted author version posted online: 01 Jun 2021, Published online: 19 Jul 2021

https://doi.org/10.1080/10618600.2021.1935971

Gen 2: Latent variable/factor analytic models

- Generalized linear latent variable models (LVMs)
  - Model residual between-species correlations using rank-reduction



Trends in Ecology & Evolution

CellPress

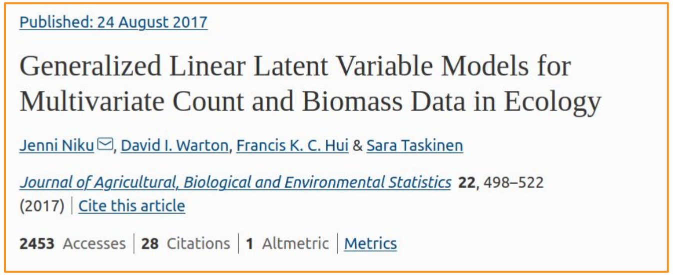
Volume 30, Issue 12, December 2015, Pages 766-779

Review

So Many Variables: Joint Modeling in Community Ecology

David I. Warton [1], F. Guillaume Blanchet [2], Robert B. O'Hara [3], Otso Ovaskainen [4, 5], Sara Taskinen [6], Steven C. Walker [2], Francis K.C. Hui [7]

https://doi.org/10.1016/j.tree.2015.09.007



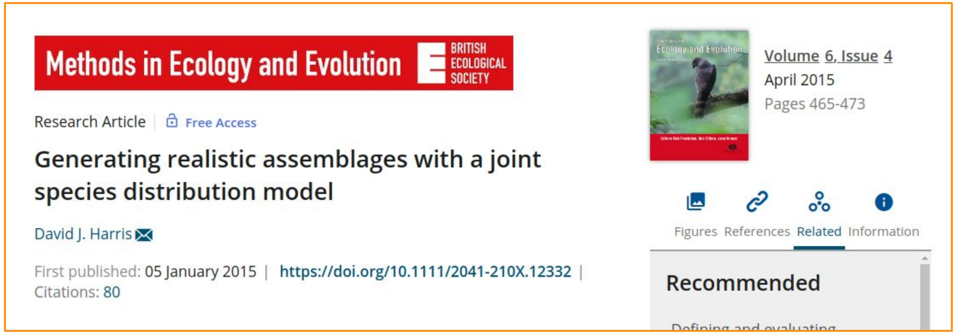
Published: 24 August 2017

Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology

Jenni Niku, David I. Warton, Francis K. C. Hui & Sara Taskinen

Journal of Agricultural, Biological and Environmental Statistics **22**, 498–522 (2017) | Cite this article

2453 Accesses | **28** Citations | **1** Altmetric | Metrics



Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Research Article | Free Access

Generating realistic assemblages with a joint species distribution model

David J. Harris

First published: 05 January 2015 | https://doi.org/10.1111/2041-210X.12332 | Citations: 80

Volume 6, Issue 4

April 2015

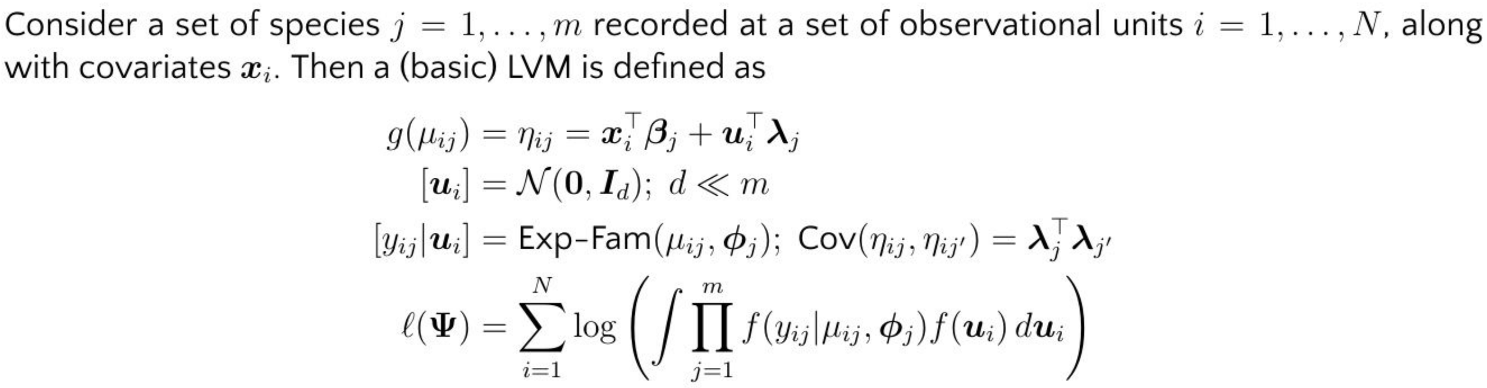
Pages 465-473

Recommended

Defining and evaluating

Gen 2: Latent variable/factor analytic models

- Generalized linear latent variable models (LVMs)
  - Model residual between-species correlations using rank-reduction

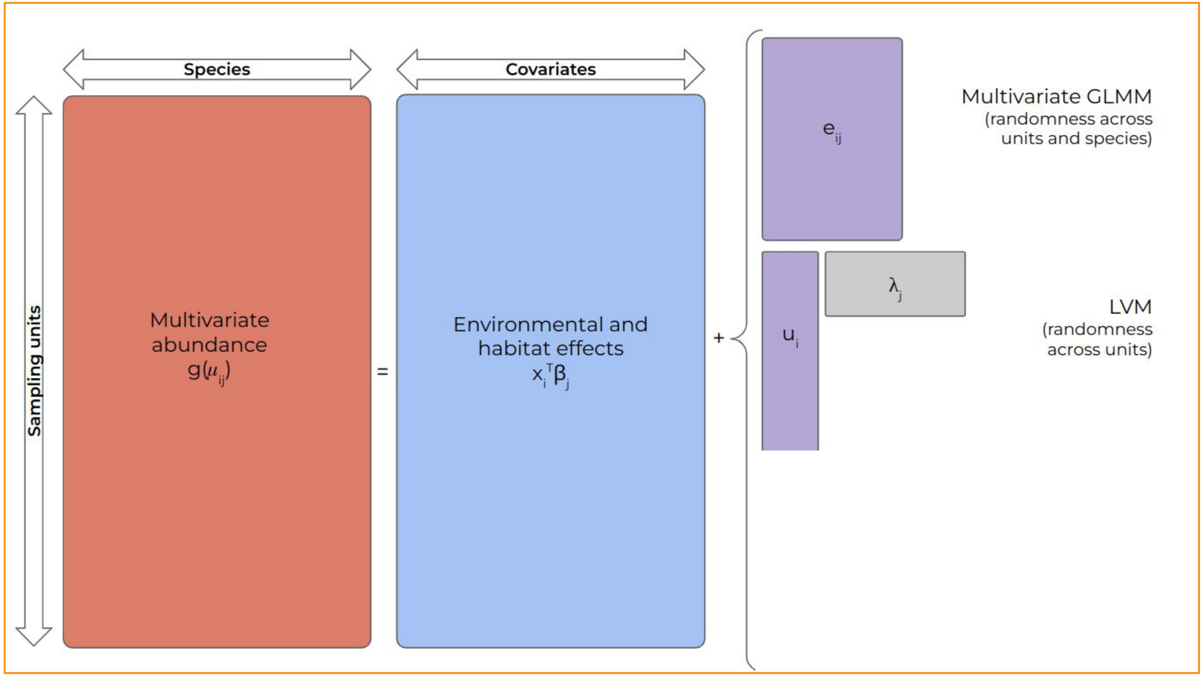


Consider a set of species $j = 1, \ldots, m$ recorded at a set of observational units $i = 1, \ldots, N$, along with covariates $\boldsymbol{x}_i$. Then a (basic) LVM is defined as

$$\begin{aligned} g(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\lambda}_j \\ [\boldsymbol{u}_i] &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d);\ d \ll m \\ [y_{ij}|\boldsymbol{u}_i] &= \text{Exp-Fam}(\mu_{ij}, \boldsymbol{\phi}_j);\ \text{Cov}(\eta_{ij}, \eta_{ij'}) = \boldsymbol{\lambda}_j^\top \boldsymbol{\lambda}_{j'} \\ \ell(\boldsymbol{\Psi}) &= \sum_{i=1}^{N} \log \left(\int \prod_{j=1}^{m} f(y_{ij}|\mu_{ij}, \boldsymbol{\phi}_j) f(\boldsymbol{u}_i)\, d\boldsymbol{u}_i \right) \end{aligned}$$

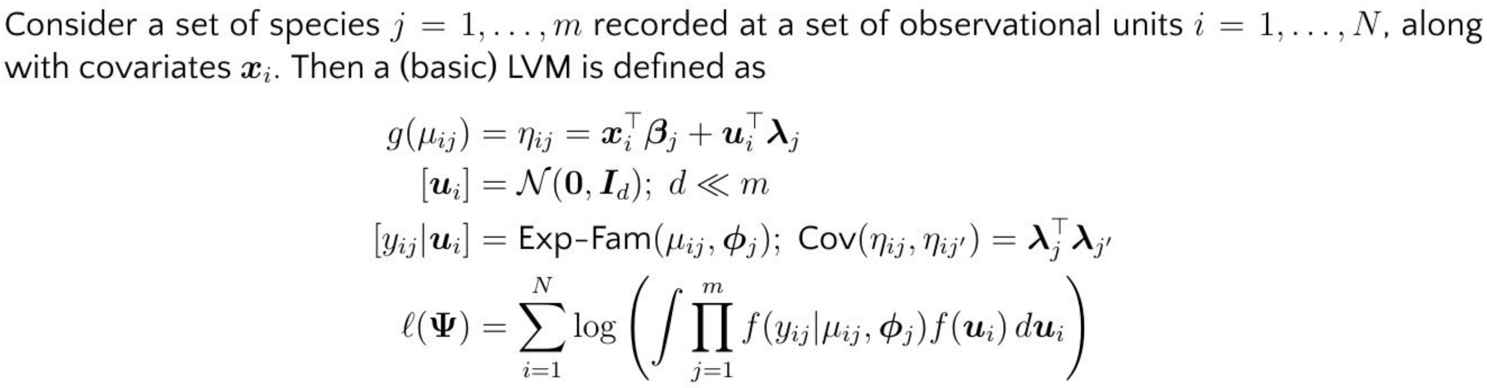
Gen 2: Latent variable/factor analytic models

- Generalized linear latent variable models (LVMs)
  - Model residual between-species correlations using rank-reduction


Species
Covariates
Sampling units
Multivariate abundance $g(\mu_{ij})$
=
Environmental and habitat effects $x_i^T\beta_j$
+
e_{ij}
Multivariate GLMM (randomness across units and species)
u_i
λ_j
LVM (randomness across units)

Gen 2: Latent variable/factor analytic models

- Generalized linear latent variable models (LVMs)
  - Model residual between-species correlations using rank-reduction
  - Less flexible than MGLMMs, but probably good enough in most scenarios?
  - Number of parameters scales as m, so can handle (a lot) more species
  - Less random effects than MGLMMs, scaling as Nd



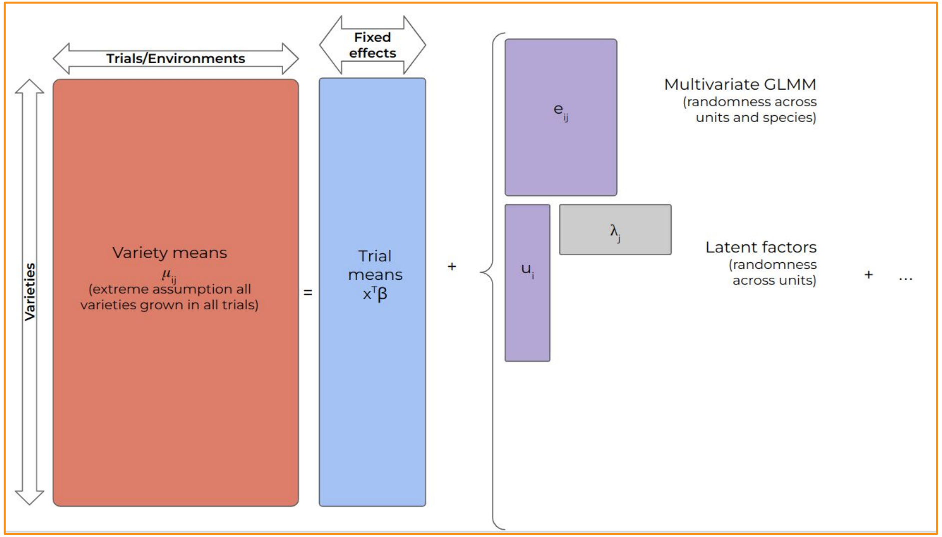
Consider a set of species $j = 1, \ldots, m$ recorded at a set of observational units $i = 1, \ldots, N$, along with covariates $\boldsymbol{x}_i$. Then a (basic) LVM is defined as

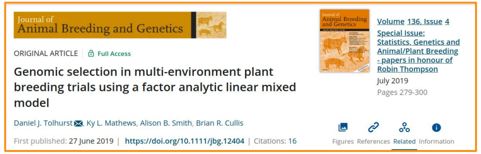
$$\begin{aligned}
g(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\lambda}_j \\
[\boldsymbol{u}_i] &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d);\ d \ll m \\
[y_{ij}|\boldsymbol{u}_i] &= \text{Exp-Fam}(\mu_{ij}, \boldsymbol{\phi}_j);\ \text{Cov}(\eta_{ij}, \eta_{ij'}) = \boldsymbol{\lambda}_j^\top \boldsymbol{\lambda}_{j'} \\
\ell(\boldsymbol{\Psi}) &= \sum_{i=1}^{N} \log \left(\int \prod_{j=1}^{m} f(y_{ij}|\mu_{ij}, \boldsymbol{\phi}_j) f(\boldsymbol{u}_i)\, d\boldsymbol{u}_i \right)
\end{aligned}$$

Gen 2: Latent variable/factor analytic models

- Generalized linear latent variable models (LVMs)
  - Model residual between-species correlations using rank-reduction

- LVMs are not new news! Examples include psychometrics, agriculture (MET)





Journal of Animal Breeding and Genetics

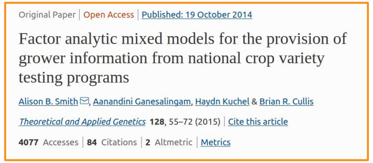
ORIGINAL ARTICLE | Full Access

Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model

Daniel J. Tolhurst, Ky L. Mathews, Alison B. Smith, Brian R. Cullis

First published: 27 June 2019 | https://doi.org/10.1111/jbg.12404 | Citations: 16

Volume 136, Issue 4
Special Issue: Statistics, Genetics and Animal/Plant Breeding - papers in honour of Robin Thompson
July 2019
Pages 279-300



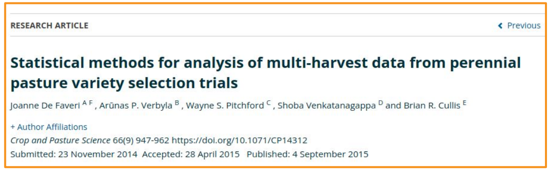
Original Paper | Open Access | Published: 19 October 2014

Factor analytic mixed models for the provision of grower information from national crop variety testing programs

Alison B. Smith, Aanandini Ganesalingam, Haydn Kuchel & Brian R. Cullis

Theoretical and Applied Genetics **128**, 55–72 (2015) | Cite this article

4077 Accesses | **84** Citations | **2** Altmetric | Metrics



RESEARCH ARTICLE ‹ Previous

Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials

Joanne De Faveri [A F], Arūnas P. Verbyla [B], Wayne S. Pitchford [C], Shoba Venkatanagappa [D] and Brian R. Cullis [E]

+ Author Affiliations

Crop and Pasture Science 66(9) 947-962 https://doi.org/10.1071/CP14312

Submitted: 23 November 2014 Accepted: 28 April 2015 Published: 4 September 2015

Gen 2: Latent variable/factor analytic models

- LVMs are not new news, but they took off in ecology!

Gen 2: Latent variable/factor analytic models

- LVMs are not new news, but they took off in ecology!
  - Model-based unconstrained/partial/concurrent ordination, when d is small

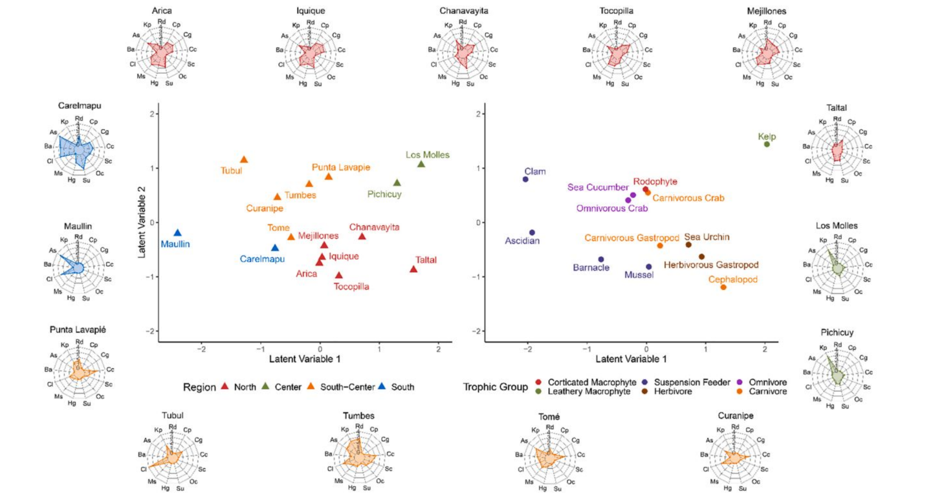
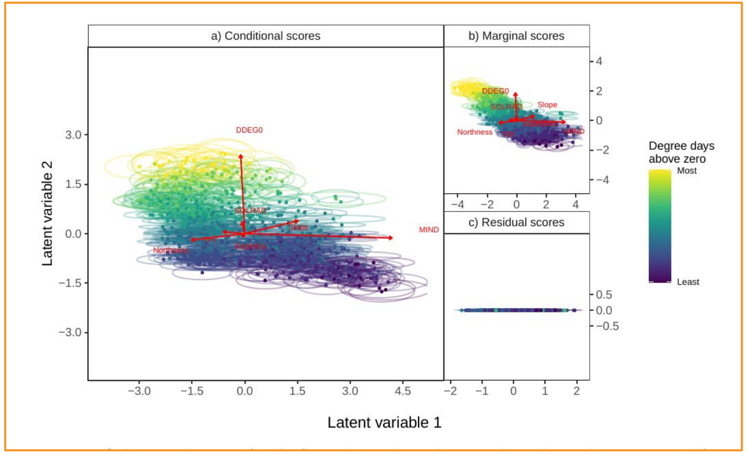


Fig. 3. Model-based unconstrained ordination of the interannual mean catch of all fishing coves for all main taxa according to a functional taxonomy combining class and trophic group. We used a Tweedie distribution function with a log link function to model biomass responses and included random row effects



$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{z}_i^\top \boldsymbol{\lambda}_j - \frac{1}{2}\boldsymbol{z}_j^\top \boldsymbol{D}_j \boldsymbol{z}_i$$

$$\boldsymbol{z}_i = \boldsymbol{C}^\top \boldsymbol{x}_{\mathrm{lv},i} + \boldsymbol{u}_i$$

$$[\boldsymbol{u}_i] = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d);\ d \ll m$$

$$[y_{ij}|\boldsymbol{u}_i] = \text{Exp-Fam}(\mu_{ij}, \phi_j)$$

Gen 2: Latent variable/factor analytic models

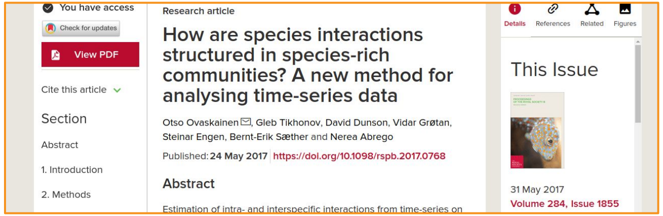
- LVMs are not new news, but they took off in ecology!
    - Model-based unconstrained/partial/concurrent ordination, when d is small
    - Latent variables interpreted as unobserved environmental predictors
        - Neat interpretation but practically not very useful

Gen 2: Latent variable/factor analytic models

- LVMs are not new news, but they took off in ecology!
  - Model-based unconstrained/partial/concurrent ordination, when d is small
  - Latent variables interpreted as unobserved environmental predictors
    - Neat interpretation but practically not very useful
  - Rank-reduction concept used in other community ecology contexts
    - Vector autoregressive models; community-level drivers/regulators

Species $j = 1, \ldots, m$ at time $t = 1, \ldots, T$

$$
\begin{aligned}
\log(\mu_{tj}) &= \boldsymbol{x}_t^\top \boldsymbol{\beta}_j + (\boldsymbol{Q}\boldsymbol{c}_j + \boldsymbol{d}_j)^\top \log(\boldsymbol{\mu}_{t-1}) + \boldsymbol{u}_t^\top \boldsymbol{\lambda}_j + \delta_j \\
\dim(\boldsymbol{Q}) &= m \times q;\ \dim(\boldsymbol{c}_j) = q \times 1;\ q \ll m \\
\boldsymbol{d}_j &= (0, 0, \ldots, 0, d_j, 0, \ldots, 0) \\
[\boldsymbol{u}_t] &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d);\ d \ll m \\
[\delta_j] &= \mathcal{N}(0, \sigma^2)
\end{aligned}
$$



Research article

How are species interactions structured in species-rich communities? A new method for analysing time-series data

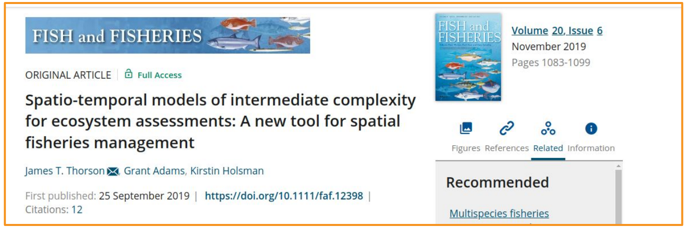
Otso Ovaskainen, Gleb Tikhonov, David Dunson, Vidar Grøtan, Steinar Engen, Bernt-Erik Sæther and Nerea Abrego

Published: 24 May 2017 https://doi.org/10.1098/rspb.2017.0768

Abstract

Estimation of intra- and interspecific interactions from time-series on

31 May 2017
Volume 284, Issue 1855



ORIGINAL ARTICLE

Spatio-temporal models of intermediate complexity for ecosystem assessments: A new tool for spatial fisheries management

James T. Thorson, Grant Adams, Kirstin Holsman

First published: 25 September 2019 | https://doi.org/10.1111/faf.12398 | Citations: 12

Volume 20, Issue 6
November 2019
Pages 1083-1099

Recommended

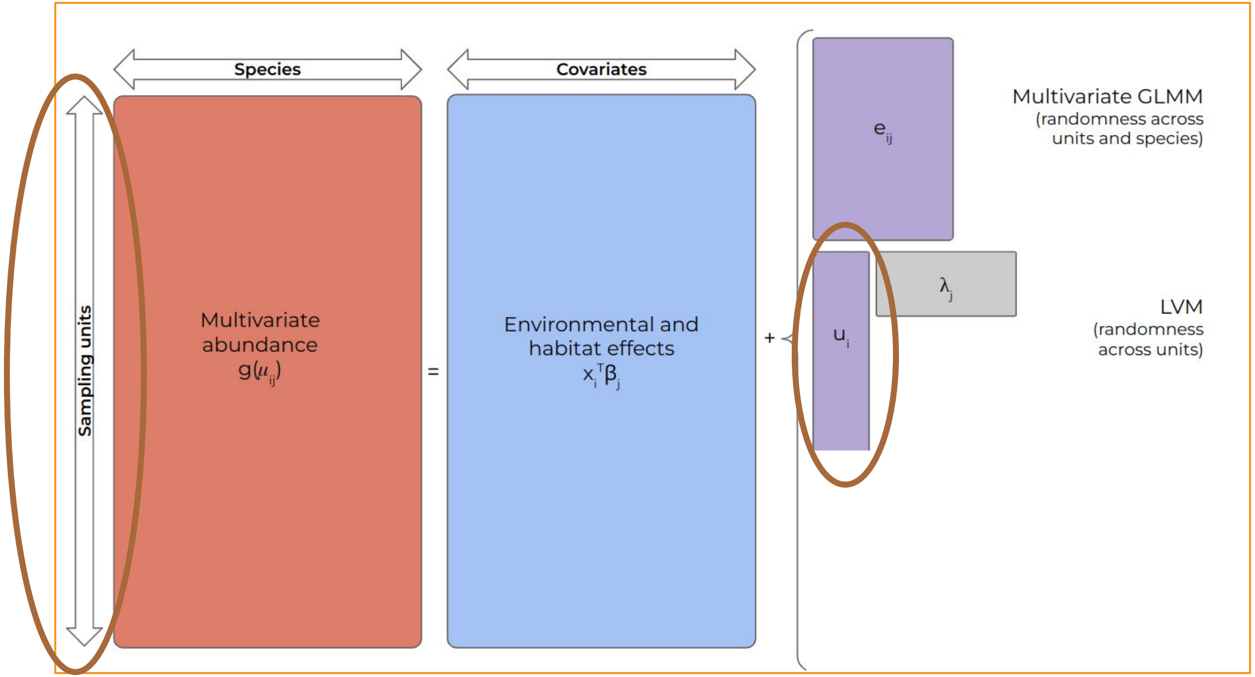
Multispecies fisheries

Gen 2+: LVMs with all the extras

- Current JSDM paradigm
  - Make LVMs more flexible and/or computationally more scalable

Gen 2+: LVMs with all the extras

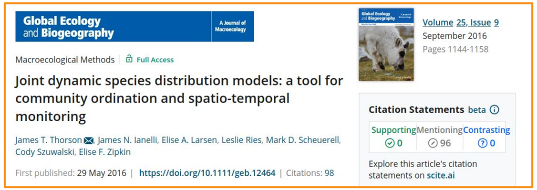
- Example 1: Spatio-temporal LVMs



Species
Covariates
Sampling units
Multivariate abundance $g(\mu_{ij})$
=
Environmental and habitat effects $x_i^T\beta_j$
+
e_{ij}
Multivariate GLMM (randomness across units and species)
u_i
λ_j
LVM (randomness across units)

Gen 2+: LVMs with all the extras

- Example 1: Spatio-temporal LVMs
  - Many flavours e.g., tensor-product or additive LVs, dynamic loadings
  - Faster approximations/algorithms e.g., LVs + SPDE/NNGP/GPP



Global Ecology and Biogeography | A Journal of Macroecology

Macroecological Methods | Full Access

Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring

James T. Thorson, James N. Ianelli, Elise A. Larsen, Leslie Ries, Mark D. Scheuerell, Cody Szuwalski, Elise F. Zipkin

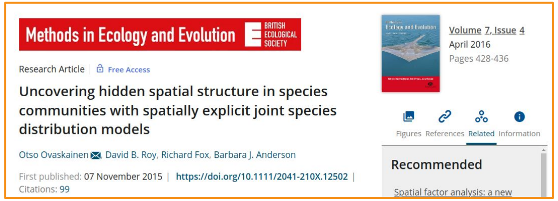
First published: 29 May 2016 | https://doi.org/10.1111/geb.12464 | Citations: 98

Volume 25, Issue 9
September 2016
Pages 1144-1158

Citation Statements beta

Supporting	Mentioning	Contrasting
0	96	0

Explore this article's citation statements on scite.ai



Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Research Article | Free Access

Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models

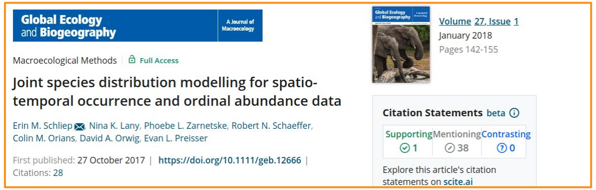
Otso Ovaskainen, David B. Roy, Richard Fox, Barbara J. Anderson

First published: 07 November 2015 | https://doi.org/10.1111/2041-210X.12502 | Citations: 99

Volume 7, Issue 4
April 2016
Pages 428-436

Recommended

Spatial factor analysis: a new



Global Ecology and Biogeography | A Journal of Macroecology

Macroecological Methods | Full Access

Joint species distribution modelling for spatio-temporal occurrence and ordinal abundance data

Erin M. Schliep, Nina K. Lany, Phoebe L. Zarnetske, Robert N. Schaeffer, Colin M. Orians, David A. Orwig, Evan L. Preisser

First published: 27 October 2017 | https://doi.org/10.1111/geb.12666 | Citations: 28

Volume 27, Issue 1
January 2018
Pages 142-155

Citation Statements beta

Supporting	Mentioning	Contrasting
1	38	0

Explore this article's citation statements on scite.ai

Consider a set of species $j = 1, \ldots, m$ recorded at a set of units $i = 1, \ldots, N$, each unit having a space-time coordinate $(\boldsymbol{s}_i, t_i)$. Then a (basic) spatio-temporal LVM is defined as

$$g\{\mu_j(\boldsymbol{s}_i, t_i)\} = \eta_j(\boldsymbol{s}_i, t_i) = \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{\beta}_j + \boldsymbol{u}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{\lambda}_j$$

$$[\boldsymbol{u}_{\cdot,k}] = [\{u_k(\boldsymbol{s}_1, t_1), \ldots, u_k(\boldsymbol{s}_N, t_N)\}] = \mathcal{N}(0, \boldsymbol{\Sigma}^{sp}_k \otimes \boldsymbol{\Sigma}^{time}_k);\ \boldsymbol{\Sigma}^{sp}_k \Rightarrow \text{Matern}(\boldsymbol{\theta}^{sp}_k), \boldsymbol{\Sigma}^{time}_k \Rightarrow \text{Matern}(\boldsymbol{\theta}^{time}_k); k = 1, \ldots, d$$

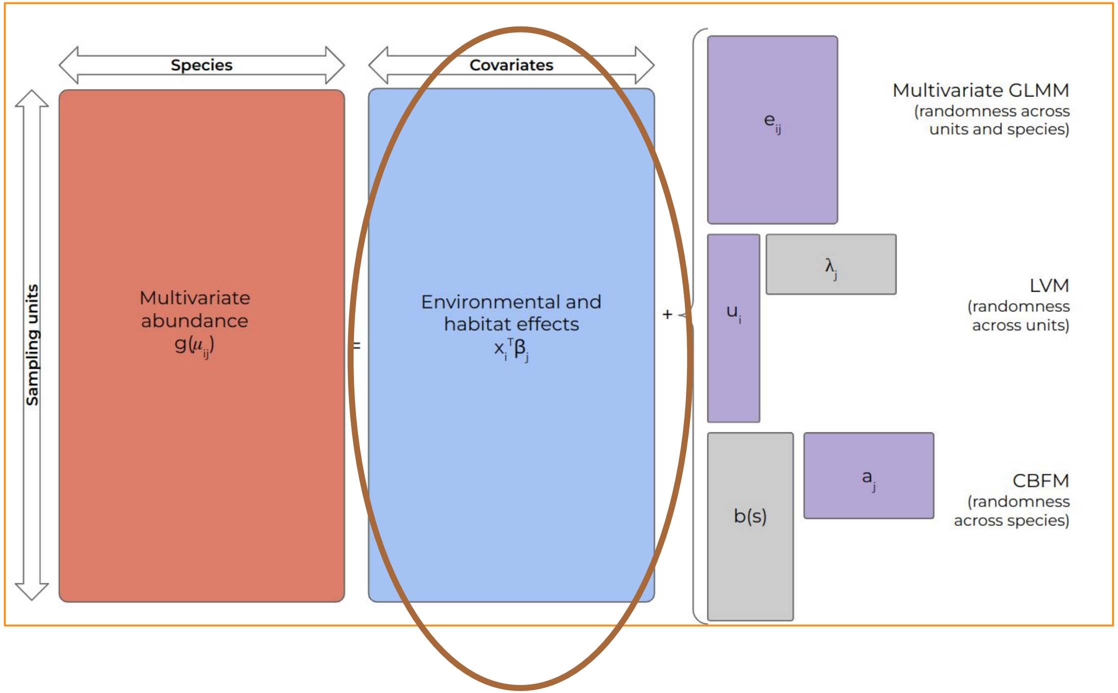
$$[y_{ij} | \boldsymbol{u}_i] = \text{Exp-Fam}(\mu_{ij}, \boldsymbol{\phi}_j)$$

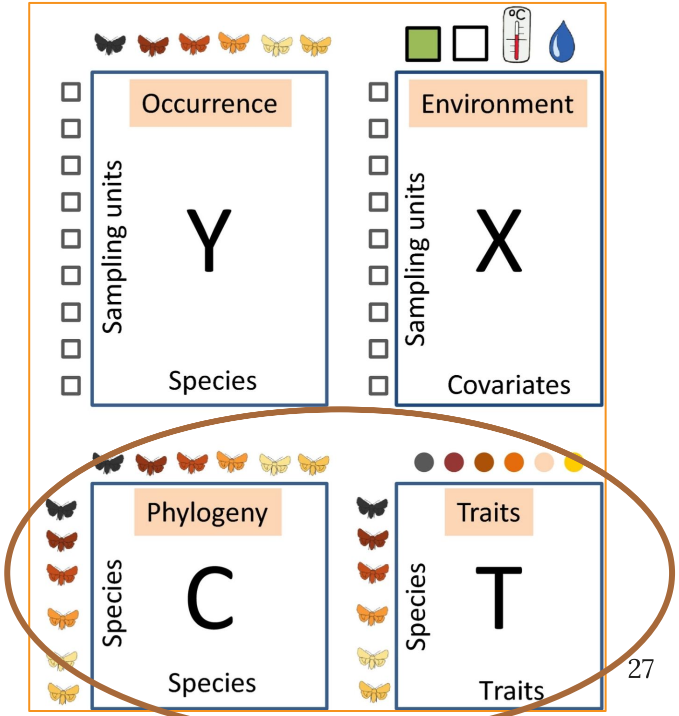
$$\ell(\boldsymbol{\Psi}) = \log\left(\int \prod_{i=1}^N \prod_{j=1}^m f(y_j(\boldsymbol{s}_i, t_i) | \mu_j(\boldsymbol{s}_i, t_i), \boldsymbol{\phi}_j) \prod_{k=1}^d f(\boldsymbol{u}_{\cdot,k}) \prod_{k=1}^d d\boldsymbol{u}_{\cdot,k}\right)$$

Note $\text{Cov}\{\eta_j(\boldsymbol{s}, t), \eta_{j'}(\boldsymbol{s}', t')\} = \sum_{k=1}^d \lambda_{jk} \Sigma^{sp}_{k,ss'} \Sigma^{time}_{k,tt'} \lambda_{j'k}$.

Gen 2+: LVMs with all the extras

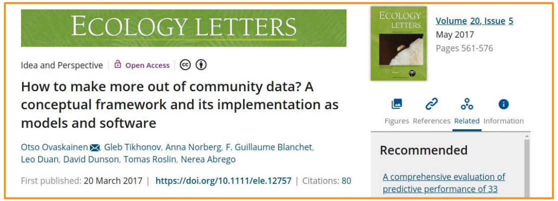
- Example 2: Borrow strength across species


Species
Covariates
Sampling units
Multivariate abundance $g(\mu_{ij})$
=
Environmental and habitat effects $x_i^T\beta_j$
+
e_{ij}
Multivariate GLMM (randomness across units and species)
u_i
λ_j
LVM (randomness across units)
b(s)
a_j
CBFM (randomness across species)



Gen 2+: LVMs with all the extras

- Example 2: Borrow strength across species
  - Traits mediate species mean responses to environment (“fourth-corner” models)
  - Phylogeny drives (dis)similarity in response to environment (phylogenetic LVMs)



ECOLOGY LETTERS

Volume 20, Issue 5
May 2017
Pages 561-576

Idea and Perspective | Open Access

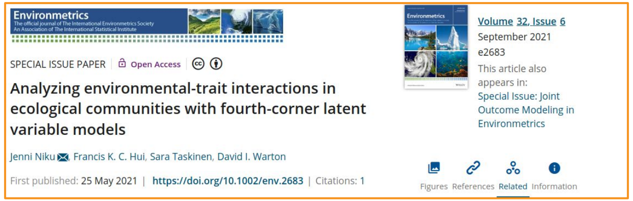
How to make more out of community data? A conceptual framework and its implementation as models and software

Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, Nerea Abrego

First published: 20 March 2017 | https://doi.org/10.1111/ele.12757 | Citations: 80

Recommended

A comprehensive evaluation of predictive performance of 33



Environmetrics
The official journal of The International Environmetrics Society
An Association of The International Statistical Institute

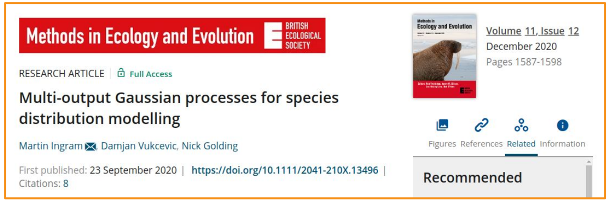
Volume 32, Issue 6
September 2021
e2683
This article also appears in:
Special Issue: Joint Outcome Modeling in Environmetrics

SPECIAL ISSUE PAPER | Open Access

Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models

Jenni Niku, Francis K. C. Hui, Sara Taskinen, David I. Warton

First published: 25 May 2021 | https://doi.org/10.1002/env.2683 | Citations: 1



Methods in Ecology and Evolution
BRITISH ECOLOGICAL SOCIETY

Volume 11, Issue 12
December 2020
Pages 1587-1598

RESEARCH ARTICLE | Full Access

Multi-output Gaussian processes for species distribution modelling

Martin Ingram, Damjan Vukcevic, Nick Golding

First published: 23 September 2020 | https://doi.org/10.1111/2041-210X.13496 | Citations: 8

Recommended

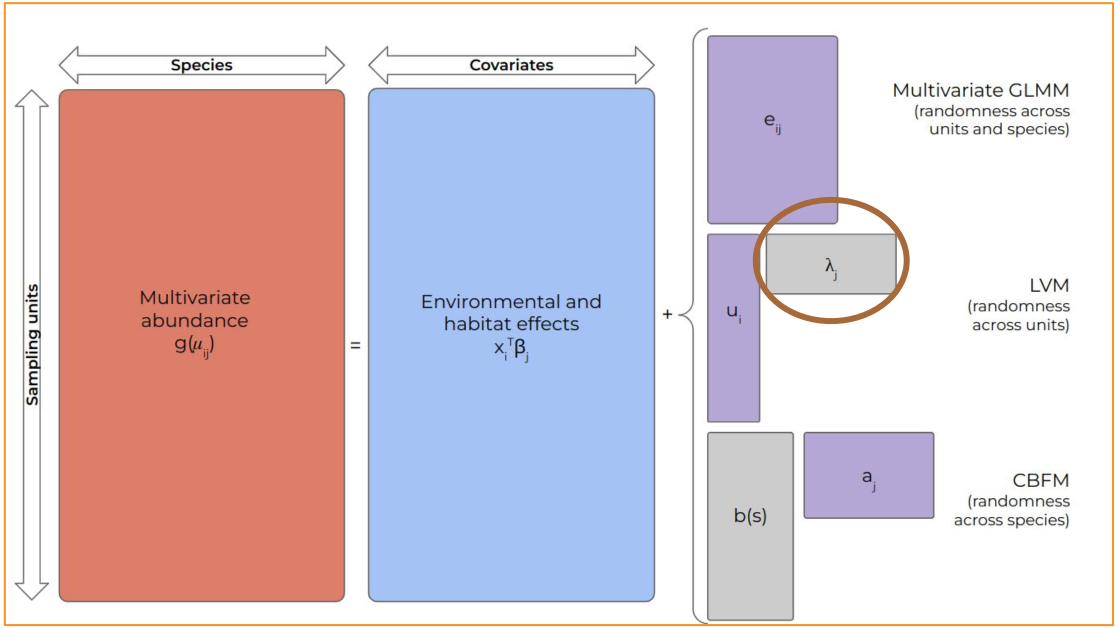
Consider a set of species $j = 1, \ldots, m$ recorded at a set of observational units $i = 1, \ldots, N$, along with a set of p covariates $\boldsymbol{x}_i$, an $m \times t$ trait matrix $\boldsymbol{T}$, and phylogenetic correlation matrix $\boldsymbol{C}$. Then a (basic) hierarchical LVM is defined as

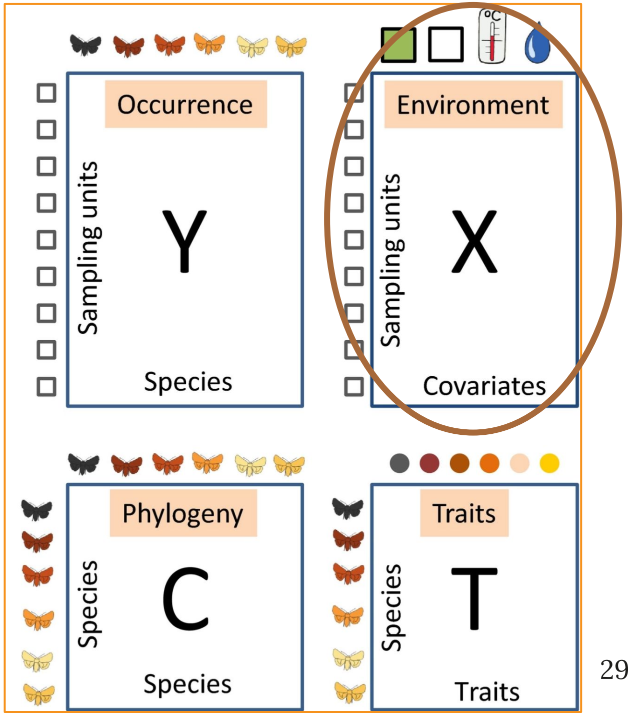
$$
\begin{aligned}
g(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\lambda}_j \\
[(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)] &= \mathcal{N}\{\text{vec}(\boldsymbol{K}\boldsymbol{T}^\top), \boldsymbol{V} \otimes (\rho \boldsymbol{C} + (1-\rho)\boldsymbol{I}_m)\} \\
\dim(\boldsymbol{K}) &= p \times t;\; t < p;\; \dim(\boldsymbol{V}) = p \times p \\
[\boldsymbol{u}_i] &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d);\; d \ll m \\
[y_{ij} | \boldsymbol{u}_i] &= \text{Exp-Fam}(\mu_{ij}, \boldsymbol{\phi}_j) \\
\ell(\boldsymbol{\Psi}) &= \log\left(\int \prod_{i=1}^{N} \prod_{j=1}^{m} f(y_{ij} | \mu_{ij}, \boldsymbol{\phi}_j) \prod_{i=1}^{N} f(\boldsymbol{u}_i) \prod_{j=1}^{m} f(\boldsymbol{\beta}_j) \prod_{i=1}^{N} d\boldsymbol{u}_i \prod_{j=1}^{m} d\boldsymbol{\beta}_j \right)
\end{aligned}
$$

Note $\text{Cov}(\eta_{ij}, \eta_{i'j'}) = \rho C_{j,j'} \boldsymbol{x}_i^\top \boldsymbol{V} \boldsymbol{x}_{i'} + \boldsymbol{\lambda}_j^\top \boldsymbol{\lambda}_{j'}$ for $j \neq j'$.

Gen 2+: LVMs with all the extras

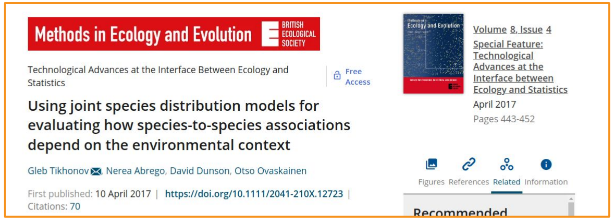
- Example 3: Borrow strength across species (in the loadings)


Species
Covariates
Sampling units
Multivariate abundance $g(\mu_{ij})$
=
Environmental and habitat effects $x_i^T\beta_j$
+
e_{ij}
Multivariate GLMM (randomness across units and species)
u_i
λ_j
LVM (randomness across units)
b(s)
a_j
CBFM (randomness across species)



Gen 2+: LVMs with all the extras

- Example 3: Borrow strength across species (in the loadings)
  - Clustering process on the loadings matrix (archetypal species associations)
  - Regress loadings against measured covariates (environment dependent associations)


Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Technological Advances at the Interface Between Ecology and Statistics

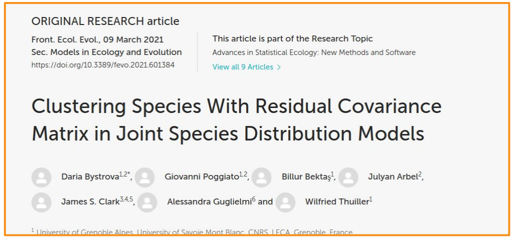
Free Access

Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context

Gleb Tikhonov, Nerea Abrego, David Dunson, Otso Ovaskainen

First published: 10 April 2017 | https://doi.org/10.1111/2041-210X.12723 | Citations: 70

Volume 8, Issue 4
Special Feature: Technological Advances at the Interface between Ecology and Statistics
April 2017
Pages 443-452


ORIGINAL RESEARCH article

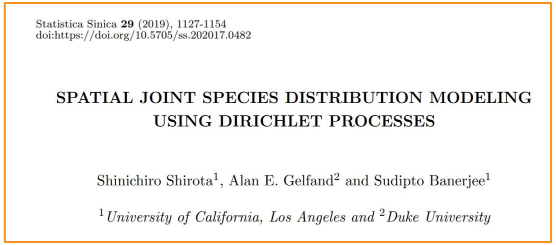
Front. Ecol. Evol., 09 March 2021
Sec. Models in Ecology and Evolution
https://doi.org/10.3389/fevo.2021.601384

This article is part of the Research Topic
Advances in Statistical Ecology: New Methods and Software
View all 9 Articles

Clustering Species With Residual Covariance Matrix in Joint Species Distribution Models

Daria Bystrova[1,2*], Giovanni Poggiato[1,2], Billur Bektaş[1], Julyan Arbel[2], James S. Clark[3,4,5], Alessandra Guglielmi[6] and Wilfried Thuiller[1]

[1] University of Grenoble Alpes, University of Savoie Mont Blanc, CNRS, LECA, Grenoble, France


Statistica Sinica **29** (2019), 1127-1154
doi:https://doi.org/10.5705/ss.202017.0482

SPATIAL JOINT SPECIES DISTRIBUTION MODELING USING DIRICHLET PROCESSES

Shinichiro Shirota[1], Alan E. Gelfand[2] and Sudipto Banerjee[1]

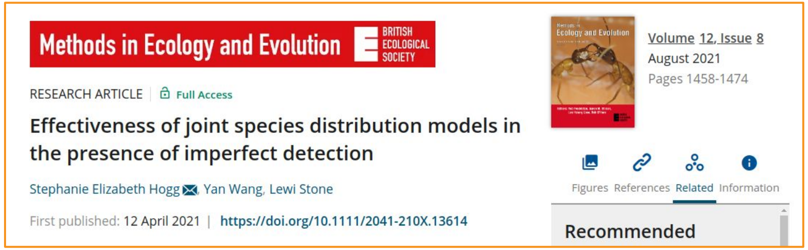
[1]*University of California, Los Angeles and* [2]*Duke University*

Consider a set of species $j = 1, \ldots, m$ recorded at a set of observational units $i = 1, \ldots, N$, along with covariates $\boldsymbol{x}_i$. Then a (basic) loading-clustered LVM is defined as

$$\begin{aligned}
g(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{Z}^\top \boldsymbol{q}(\boldsymbol{k}_j) \\
\dim(\boldsymbol{Z}) &= r \times d;\ r \gg d;\ \dim\{\boldsymbol{q}(\boldsymbol{k}_j)\} = N \times 1 \\
[\boldsymbol{k}_j] &= \mathcal{DP}(\boldsymbol{\alpha}, \{1, 2, \ldots, r\}); \\
[\boldsymbol{z}_{\cdot l}] &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{W});\ l = 1, \ldots, r \\
[\boldsymbol{u}_i] &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d);\ d \ll m \\
[y_{ij}|\boldsymbol{u}_i] &= \text{Exp-Fam}(\mu_{ij}, \boldsymbol{\phi}_j) \\
\ell(\boldsymbol{\Psi}) &= \text{I've never seen anyone try to estimate this using MLE!}
\end{aligned}$$

Gen 2+: LVMs with all the extras

- There are many other extensions of LVMs, which I do not be cover/know about!



Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

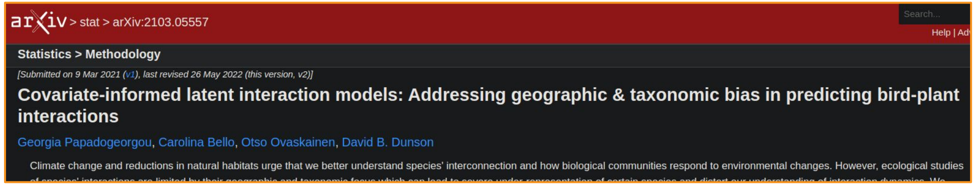
RESEARCH ARTICLE Full Access

Effectiveness of joint species distribution models in the presence of imperfect detection

Stephanie Elizabeth Hogg, Yan Wang, Lewi Stone

First published: 12 April 2021 | https://doi.org/10.1111/2041-210X.13614

Volume 12, Issue 8
August 2021
Pages 1458-1474



arXiv > stat > arXiv:2103.05557

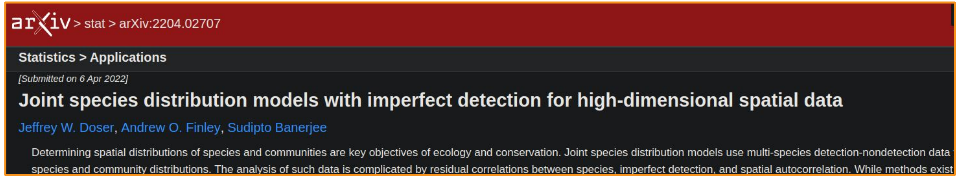
Statistics > Methodology

[Submitted on 9 Mar 2021 (v1), last revised 26 May 2022 (this version, v2)]

Covariate-informed latent interaction models: Addressing geographic & taxonomic bias in predicting bird-plant interactions

Georgia Papadogeorgou, Carolina Bello, Otso Ovaskainen, David B. Dunson

Climate change and reductions in natural habitats urge that we better understand species' interconnection and how biological communities respond to environmental changes. However, ecological studies



arXiv > stat > arXiv:2204.02707

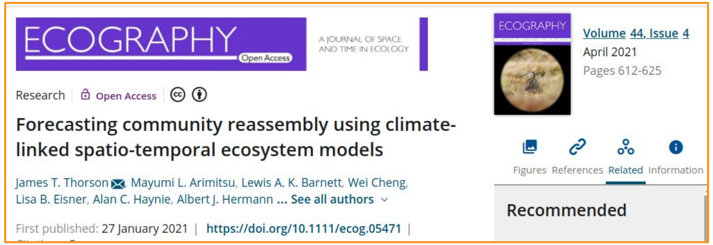
Statistics > Applications

[Submitted on 6 Apr 2022]

Joint species distribution models with imperfect detection for high-dimensional spatial data

Jeffrey W. Doser, Andrew O. Finley, Sudipto Banerjee

Determining spatial distributions of species and communities are key objectives of ecology and conservation. Joint species distribution models use multi-species detection-nondetection data species and community distributions. The analysis of such data is complicated by residual correlations between species, imperfect detection, and spatial autocorrelation. While methods exist



ECOGRAPHY Open Access A JOURNAL OF SPACE AND TIME IN ECOLOGY

Research Open Access

Forecasting community reassembly using climate-linked spatio-temporal ecosystem models

James T. Thorson, Mayumi L. Arimitsu, Lewis A. K. Barnett, Wei Cheng, Lisa B. Eisner, Alan C. Haynie, Albert J. Hermann ... See all authors

First published: 27 January 2021 | https://doi.org/10.1111/ecog.05471 |

Volume 44, Issue 4
April 2021
Pages 612-625

Some closing remarks/thoughts

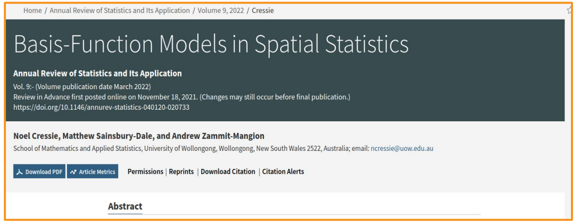
Some closing remarks/thoughts

- JSDMs is a success story of how to translate and sell statistics…
    - Targeted software + relevant interpretations/answers + methods-vs-maths gap

Some closing remarks/thoughts

- JSDMs is a success story of how to translate and sell statistics…
  - Targeted software + relevant interpretations/answers + methods-vs-maths gap

- Still many gaps in the JSDMs literature to close. Personal examples include:
  - Directional associations (structural equation modeling)?
  - Where do machine learning techniques come into this?
  - Data integration/fusion in JSDMs
  - Gen 3: Replacing latent variables with (spatio-temporal) basis functions
    - https://github.com/fhui28/CBFM



Home / Annual Review of Statistics and Its Application / Volume 9, 2022 / Cressie

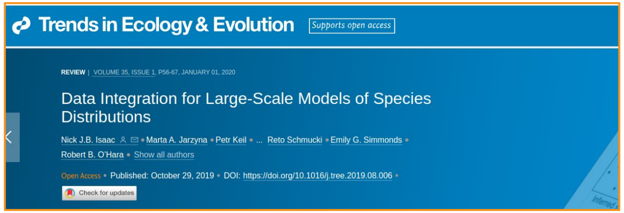
Basis-Function Models in Spatial Statistics

Annual Review of Statistics and Its Application

Vol. 9:- (Volume publication date March 2022)
Review in Advance first posted online on November 18, 2021. (Changes may still occur before final publication.)
https://doi.org/10.1146/annurev-statistics-040120-020733

Noel Cressie, Matthew Sainsbury-Dale, and Andrew Zammit-Mangion
School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, New South Wales 2522, Australia; email: ncressie@uow.edu.au

Download PDF
Article Metrics
Permissions | Reprints | Download Citation | Citation Alerts
Abstract



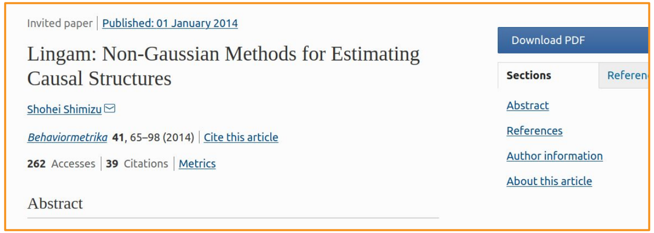
Trends in Ecology & Evolution Supports open access

REVIEW | VOLUME 35, ISSUE 1, P56-67, JANUARY 01, 2020

Data Integration for Large-Scale Models of Species Distributions

Nick J.B. Isaac • Marta A. Jarzyna • Petr Keil • ... Reto Schmucki • Emily G. Simmonds • Robert B. O'Hara • Show all authors

Open Access • Published: October 29, 2019 • DOI: https://doi.org/10.1016/j.tree.2019.08.006 •



Invited paper | Published: 01 January 2014

Lingam: Non-Gaussian Methods for Estimating Causal Structures

Shohei Shimizu

Behaviormetrika **41**, 65–98 (2014) | Cite this article

262 Accesses | **39** Citations | Metrics

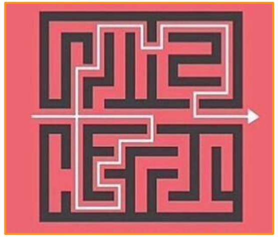
Abstract

Some closing remarks/thoughts

- JSDMs is a success story of how to translate and sell statistics...
    - Targeted software + relevant interpretations/answers + methods-vs-maths gap

- Still many gaps in the JSDMs literature to close. Personal examples include:
    - Directional associations (structural equation modeling)?
    - Where do machine learning techniques come into this?
    - Data integration/fusion in JSDMs
    - Gen 3: Replacing latent variables with (spatio-temporal) basis functions
        - https://github.com/fhui28/CBFM

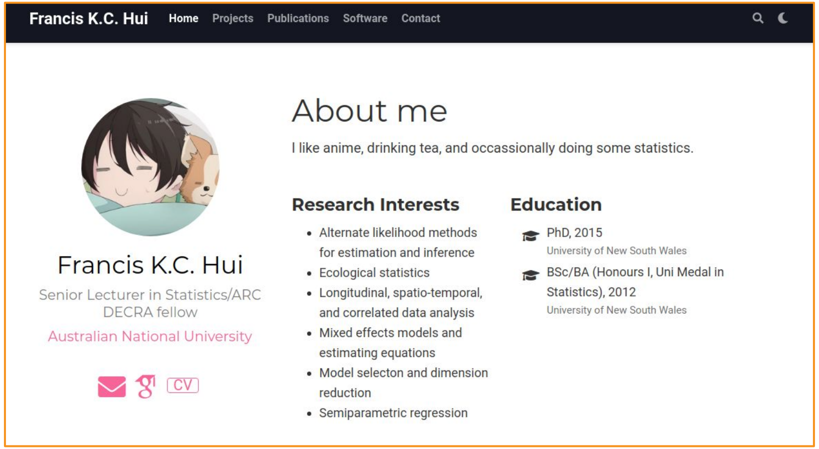
- JSDMs is not the be-all and end-all
    - E.g., Stacked SDMs are still a powerful statistical approach
    - Do not throw the kitchen sink at something that does not need it



Thank you for listening!

Any *questions*?

- francis.hui@anu.edu.au
- https://francishui.netlify.app/



Francis K.C. Hui

Francis K.C. Hui

Senior Lecturer in Statistics/ARC DECRA fellow

Australian National University

About me

I like anime, drinking tea, and occassionally doing some statistics.

Research Interests

- Alternate likelihood methods for estimation and inference
- Ecological statistics
- Longitudinal, spatio-temporal, and correlated data analysis
- Mixed effects models and estimating equations
- Model selecton and dimension reduction
- Semiparametric regression

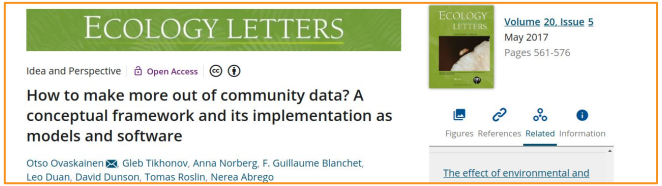
Education

- PhD, 2015
  University of New South Wales
- BSc/BA (Honours I, Uni Medal in Statistics), 2012
  University of New South Wales

Multivariate abundance data

- Some common features:
  - Multiple correlated responses (high-dimensional)
  - Non-continuous responses with evident mean-variance relationship
  - Non-linear Y-X relationships

- Other features:
  - Spatio-temporal (high-volume)
  - Multiple data sources
  - Background information



ECOLOGY LETTERS

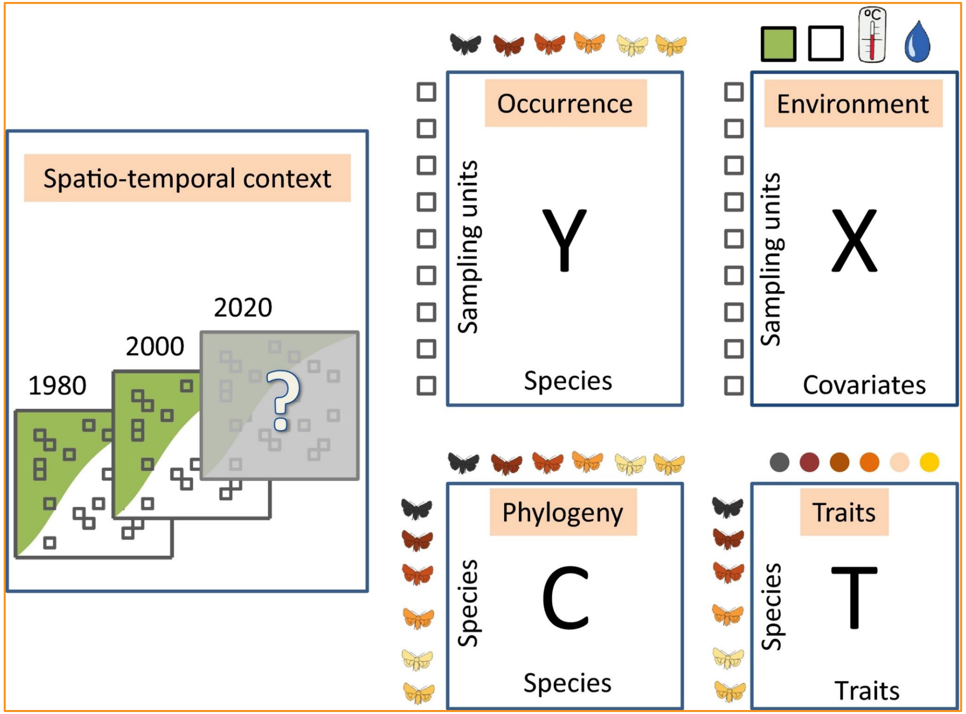
Idea and Perspective | Open Access

How to make more out of community data? A conceptual framework and its implementation as models and software

Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, Nerea Abrego

Volume 20, Issue 5
May 2017
Pages 561-576

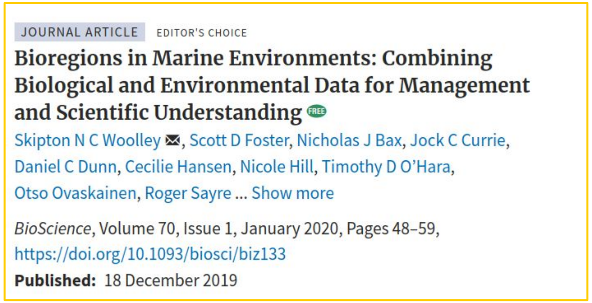
The effect of environmental and



Question/s of interests

- Depends on the data you have:
  - (a) is a multivariate prediction problem
  - (b) -> how is Y and X related?
  - (c) -> how are the columns of Y related?
  - (d) + (e) -> how do T & C mediate/drive the Y-X relationship?

- Some other applications:
  - Model-based ordination
  - Bioregionalization



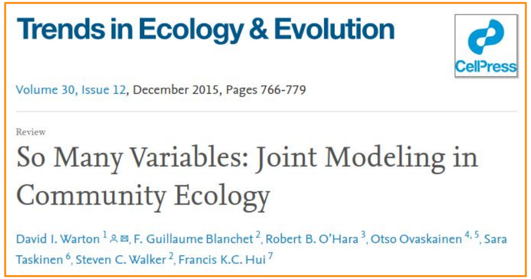
JOURNAL ARTICLE EDITOR'S CHOICE

Bioregions in Marine Environments: Combining Biological and Environmental Data for Management and Scientific Understanding FREE

Skipton N C Woolley ✉, Scott D Foster, Nicholas J Bax, Jock C Currie, Daniel C Dunn, Cecilie Hansen, Nicole Hill, Timothy D O'Hara, Otso Ovaskainen, Roger Sayre ... Show more

BioScience, Volume 70, Issue 1, January 2020, Pages 48–59, https://doi.org/10.1093/biosci/biz133

Published: 18 December 2019



Trends in Ecology & Evolution

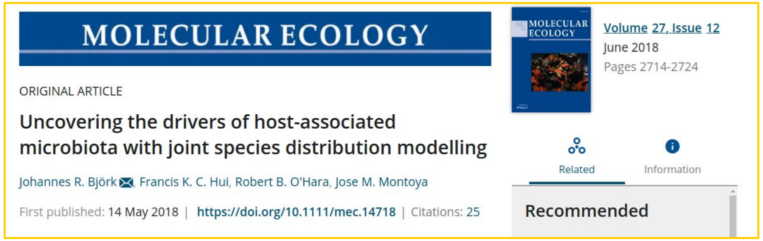
CellPress

Volume 30, Issue 12, December 2015, Pages 766-779

Review

So Many Variables: Joint Modeling in Community Ecology

David I. Warton[1] ✉, F. Guillaume Blanchet[2], Robert B. O'Hara[3], Otso Ovaskainen[4, 5], Sara Taskinen[6], Steven C. Walker[2], Francis K.C. Hui[7]



MOLECULAR ECOLOGY

Volume 27, Issue 12
June 2018
Pages 2714-2724

ORIGINAL ARTICLE

Uncovering the drivers of host-associated microbiota with joint species distribution modelling

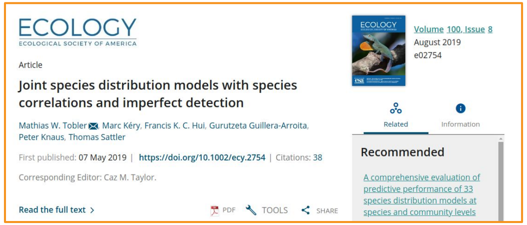
Johannes R. Björk ✉, Francis K. C. Hui, Robert B. O'Hara, Jose M. Montoya

First published: 14 May 2018 | https://doi.org/10.1111/mec.14718 | Citations: 25

Gen 2: Latent variable/factor analytic models

- Generalized linear latent variable models (LVMs)
  - Model residual between-response correlations using rank-reduction
  - Less flexible than MGLMMs, but probably good enough in most scenarios?*
  - Number of parameters scales as m, so can handle (a lot) more species*
  - Less random effects than MGLMMs, scaling as Nd; still quite challenging to fit**
    - *Choice of d remains a complicated and active topic
    - **Lots of work has been done in this space

*



ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

Article

Joint species distribution models with species correlations and imperfect detection

Mathias W. Tobler, Marc Kéry, Francis K. C. Hui, Gurutzeta Guillera-Arroita, Peter Knaus, Thomas Sattler

First published: 07 May 2019 | https://doi.org/10.1002/ecy.2754 | Citations: 38

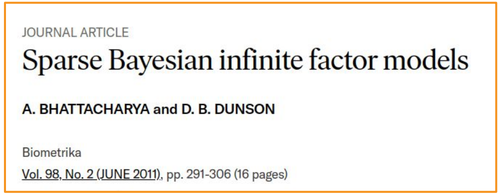
Corresponding Editor: Caz M. Taylor.

Read the full text >

Volume 100, Issue 8
August 2019
e02754

Recommended

A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels



JOURNAL ARTICLE

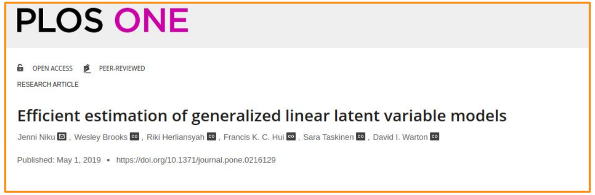
Sparse Bayesian infinite factor models

A. BHATTACHARYA and D. B. DUNSON

Biometrika

Vol. 98, No. 2 (JUNE 2011), pp. 291-306 (16 pages)

**



PLOS ONE

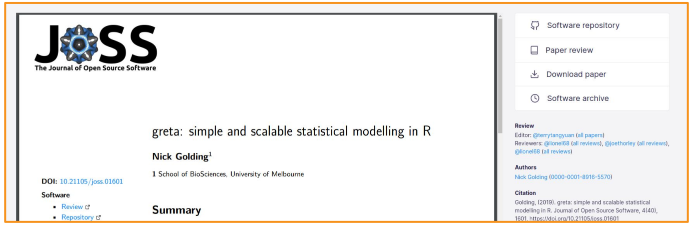
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Efficient estimation of generalized linear latent variable models

Jenni Niku, Wesley Brooks, Riki Herliansyah, Francis K. C. Hui, Sara Taskinen, David I. Warton

Published: May 1, 2019 • https://doi.org/10.1371/journal.pone.0216129



JOSS
The Journal of Open Source Software

greta: simple and scalable statistical modelling in R

Nick Golding[1]

1 School of BioSciences, University of Melbourne

DOI: 10.21105/joss.01601

Software
- Review
- Repository

Summary

Review
Editor: @terrytangyuan (all papers)
Reviewers: @lionel68 (all reviews), @joethorley (all reviews), @lionel68 (all reviews)

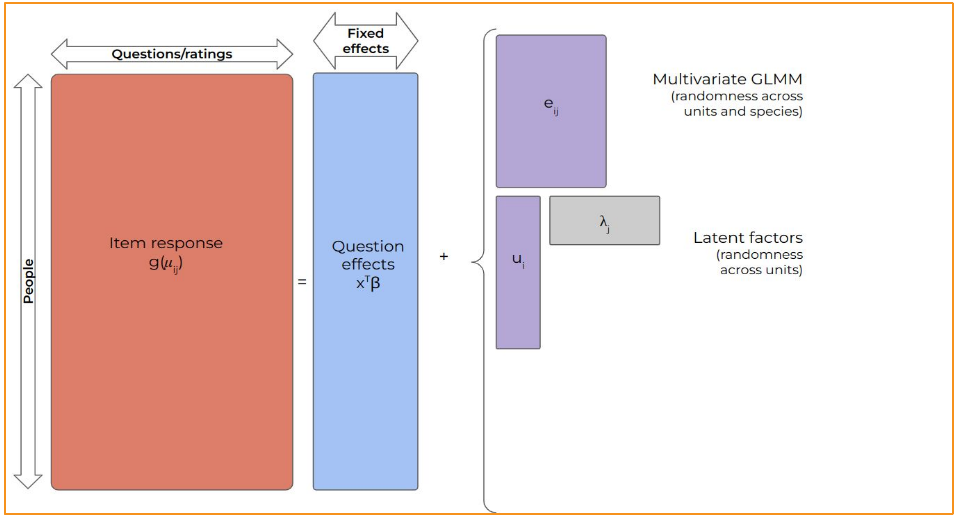
Authors
Nick Golding (0000-0001-8916-5570)

Citation
Golding, (2019). greta: simple and scalable statistical modelling in R. Journal of Open Source Software, 4(40), 1601. https://doi.org/10.21105/joss.01601

Gen 2: Latent variable/factor analytic models

- Generalized linear latent variable models (LVMs)
    - Model residual between-species correlations using rank-reduction

- LVMs are not new news! Examples include psychometrics



Questions/ratings

Fixed effects

People

Item response $g(\mu_{ij})$

=

Question effects $x^T\beta$

+

e_{ij}

Multivariate GLMM (randomness across units and species)

u_i

λ_j

Latent factors (randomness across units)

issues [read more].

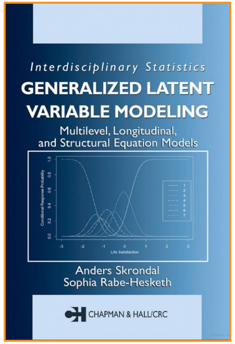
mirt: A Multidimensional Item Response Theory Package for the R Environment

R. Philip Chalmers

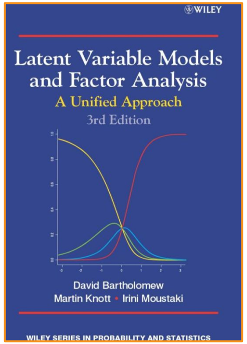
Abstract

How to Cite

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06



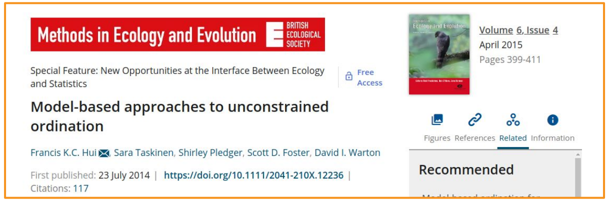
Interdisciplinary Statistics
GENERALIZED LATENT VARIABLE MODELING
Multilevel, Longitudinal, and Structural Equation Models
Anders Skrondal
Sophia Rabe-Hesketh
CHAPMAN & HALL/CRC



WILEY
Latent Variable Models and Factor Analysis
A Unified Approach
3rd Edition
David Bartholomew
Martin Knott • Irini Moustaki
WILEY SERIES IN PROBABILITY AND STATISTICS

Gen 2: Latent variable/factor analytic models

- LVMs are not new news, but they took off in ecology!
  - Model-based unconstrained and partial ordination, when d is small



Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Volume 6, Issue 4
April 2015
Pages 399-411

Special Feature: New Opportunities at the Interface Between Ecology and Statistics

Free Access

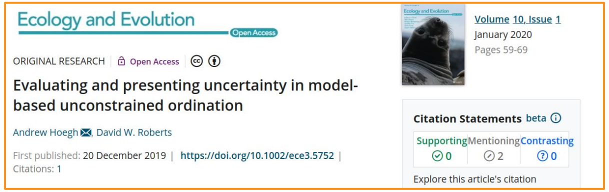
Model-based approaches to unconstrained ordination

Francis K.C. Hui, Sara Taskinen, Shirley Pledger, Scott D. Foster, David I. Warton

First published: 23 July 2014 | https://doi.org/10.1111/2041-210X.12236 |
Citations: 117

Recommended

Model-based ordination for



Ecology and Evolution Open Access

Volume 10, Issue 1
January 2020
Pages 59-69

ORIGINAL RESEARCH | Open Access

Evaluating and presenting uncertainty in model-based unconstrained ordination

Andrew Hoegh, David W. Roberts

First published: 20 December 2019 | https://doi.org/10.1002/ece3.5752 |
Citations: 1

Citation Statements beta

Supporting	Mentioning	Contrasting
0	2	0

Explore this article's citation

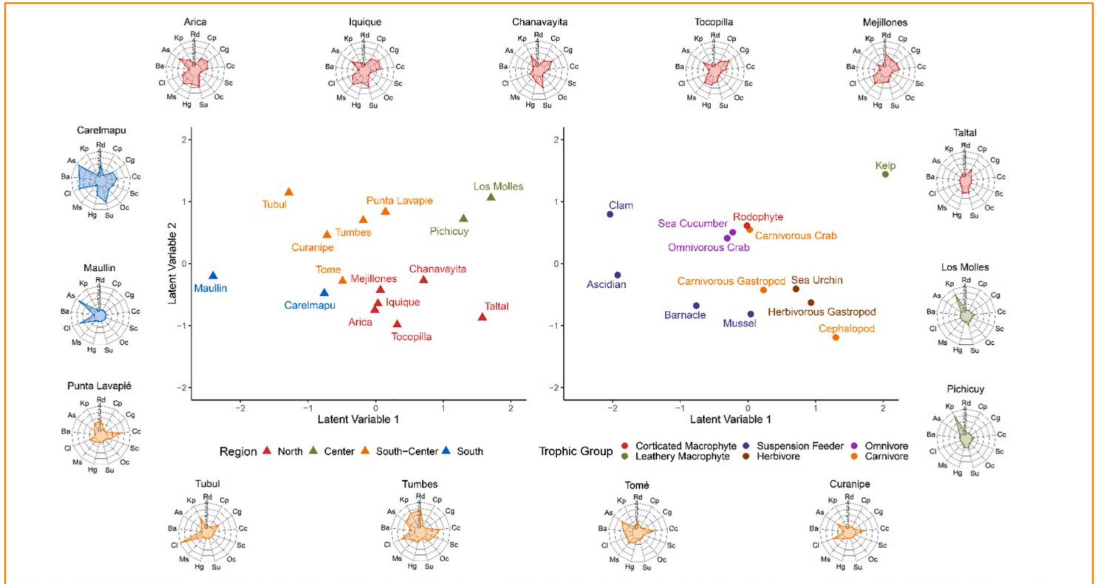
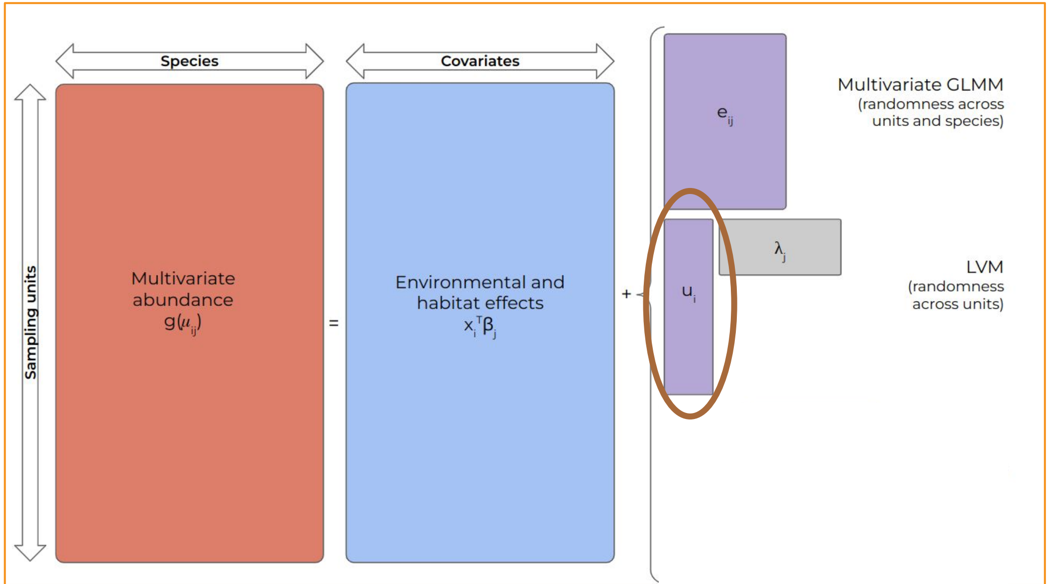


Fig. 3. **Model-based unconstrained ordination of the interannual mean catch of all fishing coves for all main taxa according to a functional taxonomy combining class and trophic group.** We used a Tweedie distribution function with a log link function to model biomass responses and included random row effects

Gen 3: ???

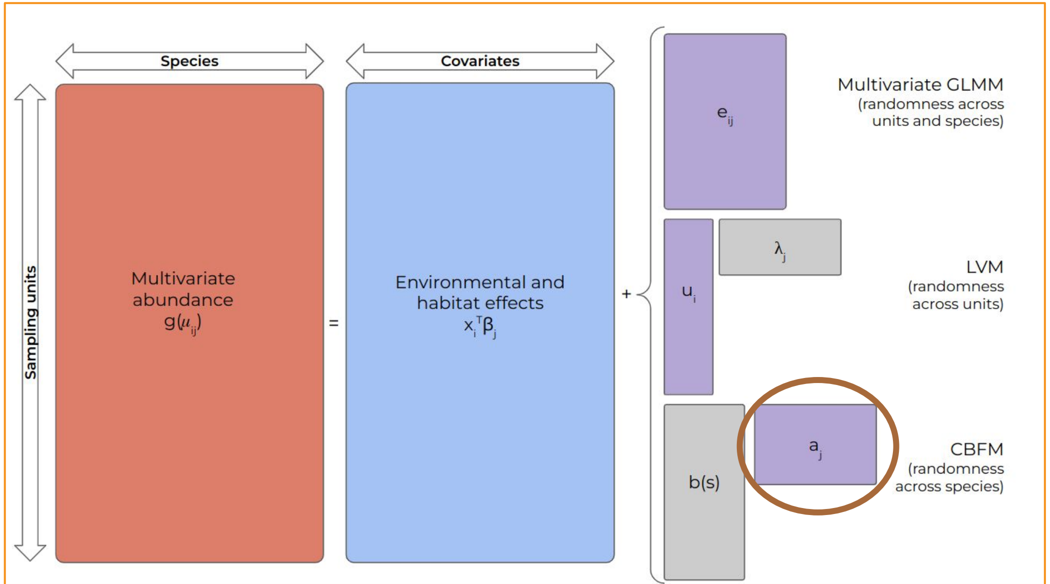
- Latent variables as an approach to JSDMs is awesome
  - But I think we are pushing the limits of their scalability/computability?


Species
Covariates
Sampling units
Multivariate abundance $g(\mu_{ij})$
=
Environmental and habitat effects $x_i^T\beta_j$
+
e_{ij}
Multivariate GLMM (randomness across units and species)
u_i
λ_j
LVM (randomness across units)

Gen 3: ???

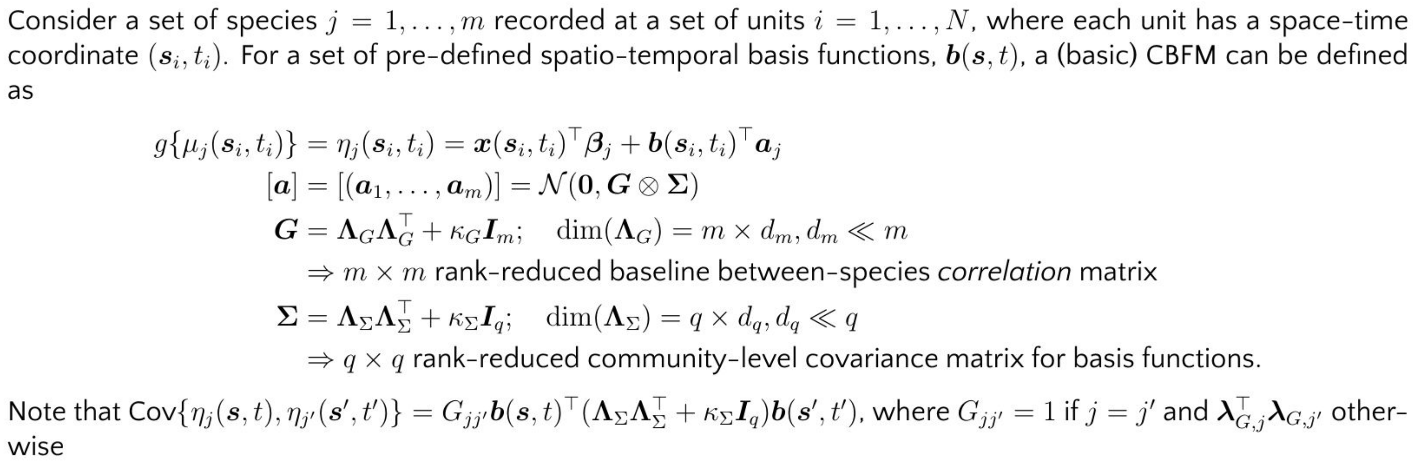
- Latent variables as an approach to JSDMs is awesome
  - But I think we are pushing the limits of their scalability/computability?

- Move the randomness from units to species -> basis functions


Species
Covariates
Sampling units
Multivariate abundance $g(\mu_{ij})$
=
Environmental and habitat effects $x_i^T\beta_j$
+
e_{ij}
Multivariate GLMM (randomness across units and species)
u_i
λ_j
LVM (randomness across units)
b(s)
a_j
CBFM (randomness across species)

Gen 3: CBFMs?

- Community-level basis function models (CBFMs) for spatio-temporal multivariate abundance data
  - Pre-defined spatio-temporal basis functions
  - https://github.com/fhui28/CBFM



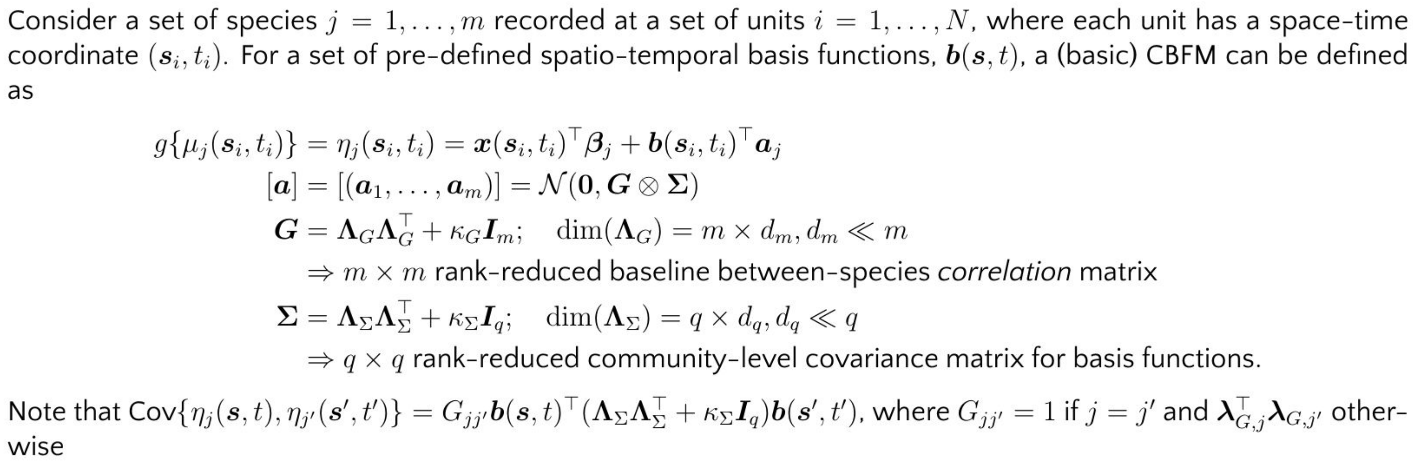
Consider a set of species $j = 1, \ldots, m$ recorded at a set of units $i = 1, \ldots, N$, where each unit has a space-time coordinate $(\boldsymbol{s}_i, t_i)$. For a set of pre-defined spatio-temporal basis functions, $\boldsymbol{b}(\boldsymbol{s}, t)$, a (basic) CBFM can be defined as

$$
\begin{aligned}
g\{\mu_j(\boldsymbol{s}_i, t_i)\} &= \eta_j(\boldsymbol{s}_i, t_i) = \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{\beta}_j + \boldsymbol{b}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{a}_j \\
[\boldsymbol{a}] &= [(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)] = \mathcal{N}(\boldsymbol{0}, \boldsymbol{G} \otimes \boldsymbol{\Sigma}) \\
\boldsymbol{G} &= \boldsymbol{\Lambda}_G \boldsymbol{\Lambda}_G^\top + \kappa_G \boldsymbol{I}_m; \quad \dim(\boldsymbol{\Lambda}_G) = m \times d_m, d_m \ll m \\
&\Rightarrow m \times m \text{ rank-reduced baseline between-species } \textit{correlation} \text{ matrix} \\
\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q; \quad \dim(\boldsymbol{\Lambda}_\Sigma) = q \times d_q, d_q \ll q \\
&\Rightarrow q \times q \text{ rank-reduced community-level covariance matrix for basis functions.}
\end{aligned}
$$

Note that Cov$\{\eta_j(\boldsymbol{s}, t), \eta_{j'}(\boldsymbol{s}', t')\} = G_{jj'} \boldsymbol{b}(\boldsymbol{s}, t)^\top (\boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q) \boldsymbol{b}(\boldsymbol{s}', t')$, where $G_{jj'} = 1$ if $j = j'$ and $\boldsymbol{\lambda}_{G,j}^\top \boldsymbol{\lambda}_{G,j'}$ otherwise

Gen 3: CBFMs?

- But why would CBFMs be faster?



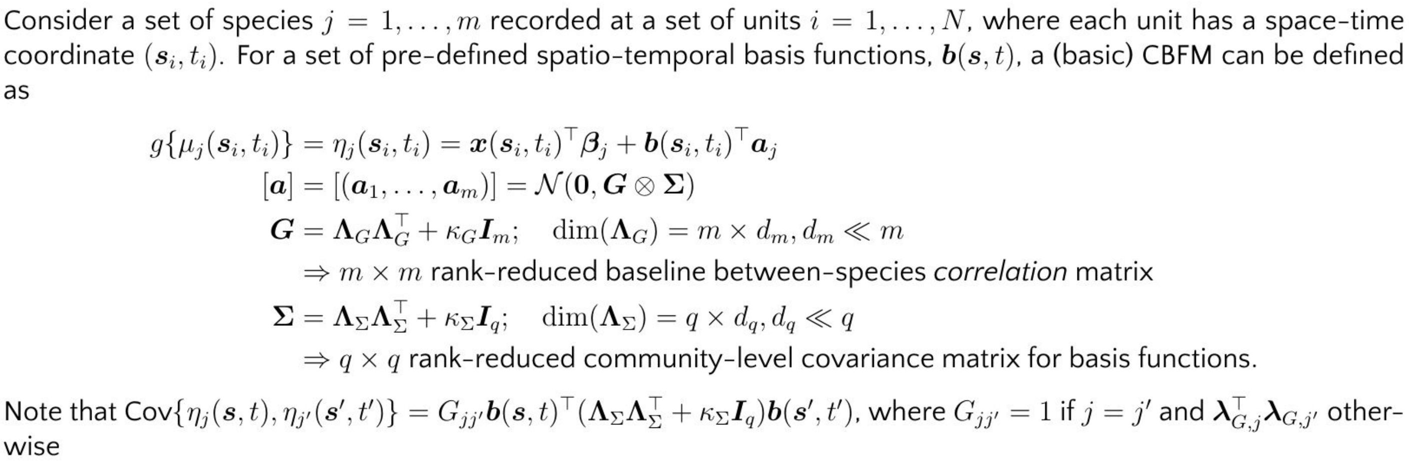
Consider a set of species $j = 1, \ldots, m$ recorded at a set of units $i = 1, \ldots, N$, where each unit has a space-time coordinate $(\boldsymbol{s}_i, t_i)$. For a set of pre-defined spatio-temporal basis functions, $\boldsymbol{b}(\boldsymbol{s}, t)$, a (basic) CBFM can be defined as

$$\begin{aligned} g\{\mu_j(\boldsymbol{s}_i, t_i)\} &= \eta_j(\boldsymbol{s}_i, t_i) = \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{\beta}_j + \boldsymbol{b}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{a}_j \\ [\boldsymbol{a}] &= [(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)] = \mathcal{N}(\boldsymbol{0}, \boldsymbol{G} \otimes \boldsymbol{\Sigma}) \\ \boldsymbol{G} &= \boldsymbol{\Lambda}_G \boldsymbol{\Lambda}_G^\top + \kappa_G \boldsymbol{I}_m; \quad \dim(\boldsymbol{\Lambda}_G) = m \times d_m, d_m \ll m \\ &\Rightarrow m \times m \text{ rank-reduced baseline between-species \textit{correlation} matrix} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q; \quad \dim(\boldsymbol{\Lambda}_\Sigma) = q \times d_q, d_q \ll q \\ &\Rightarrow q \times q \text{ rank-reduced community-level covariance matrix for basis functions.} \end{aligned}$$

Note that Cov$\{\eta_j(\boldsymbol{s}, t), \eta_{j'}(\boldsymbol{s}', t')\} = G_{jj'} \boldsymbol{b}(\boldsymbol{s}, t)^\top (\boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q) \boldsymbol{b}(\boldsymbol{s}', t')$, where $G_{jj'} = 1$ if $j = j'$ and $\boldsymbol{\lambda}_{G,j}^\top \boldsymbol{\lambda}_{G,j'}$ otherwise

Gen 3: CBFMs?

- But why would CBFMs be faster?
  - Although m may not be small, N is still larger in most modern datasets



Consider a set of species $j = 1, \ldots, m$ recorded at a set of units $i = 1, \ldots, N$, where each unit has a space-time coordinate $(\boldsymbol{s}_i, t_i)$. For a set of pre-defined spatio-temporal basis functions, $\boldsymbol{b}(\boldsymbol{s}, t)$, a (basic) CBFM can be defined as

$$g\{\mu_j(\boldsymbol{s}_i, t_i)\} = \eta_j(\boldsymbol{s}_i, t_i) = \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{\beta}_j + \boldsymbol{b}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{a}_j$$

$$[\boldsymbol{a}] = [(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)] = \mathcal{N}(\boldsymbol{0}, \boldsymbol{G} \otimes \boldsymbol{\Sigma})$$

$$\boldsymbol{G} = \boldsymbol{\Lambda}_G \boldsymbol{\Lambda}_G^\top + \kappa_G \boldsymbol{I}_m; \quad \dim(\boldsymbol{\Lambda}_G) = m \times d_m, d_m \ll m$$

$\Rightarrow m \times m$ rank-reduced baseline between-species *correlation* matrix

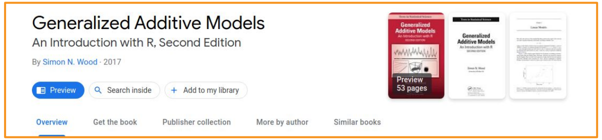
$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q; \quad \dim(\boldsymbol{\Lambda}_\Sigma) = q \times d_q, d_q \ll q$$

$\Rightarrow q \times q$ rank-reduced community-level covariance matrix for basis functions.

Note that $\text{Cov}\{\eta_j(\boldsymbol{s}, t), \eta_{j'}(\boldsymbol{s}', t')\} = G_{jj'} \boldsymbol{b}(\boldsymbol{s}, t)^\top (\boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q) \boldsymbol{b}(\boldsymbol{s}', t')$, where $G_{jj'} = 1$ if $j = j'$ and $\boldsymbol{\lambda}_{G,j}^\top \boldsymbol{\lambda}_{G,j'}$ otherwise

Gen 3: CBFMs?

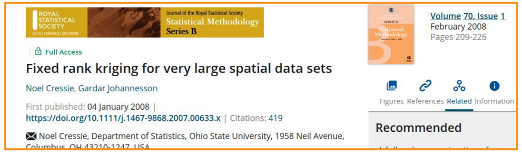
- But why would CBFMs be faster?
  - Although m may not be small, N is still larger in most modern datasets
  - It is just a big generalized additive model (GAM)!



Generalized Additive Models

An Introduction with R, Second Edition

By Simon N. Wood · 2017



Journal of the Royal Statistical Society Statistical Methodology Series B

Full Access

Fixed rank kriging for very large spatial data sets

Noel Cressie, Gardar Johannesson

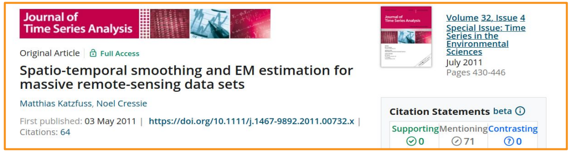
First published: 04 January 2008 | https://doi.org/10.1111/j.1467-9868.2007.00633.x | Citations: 419

✉ Noel Cressie, Department of Statistics, Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247, USA

Volume 70, Issue 1

February 2008

Pages 209-226



Journal of Time Series Analysis

Original Article | Full Access

Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets

Matthias Katzfuss, Noel Cressie

First published: 03 May 2011 | https://doi.org/10.1111/j.1467-9892.2011.00732.x | Citations: 64

Volume 32, Issue 4

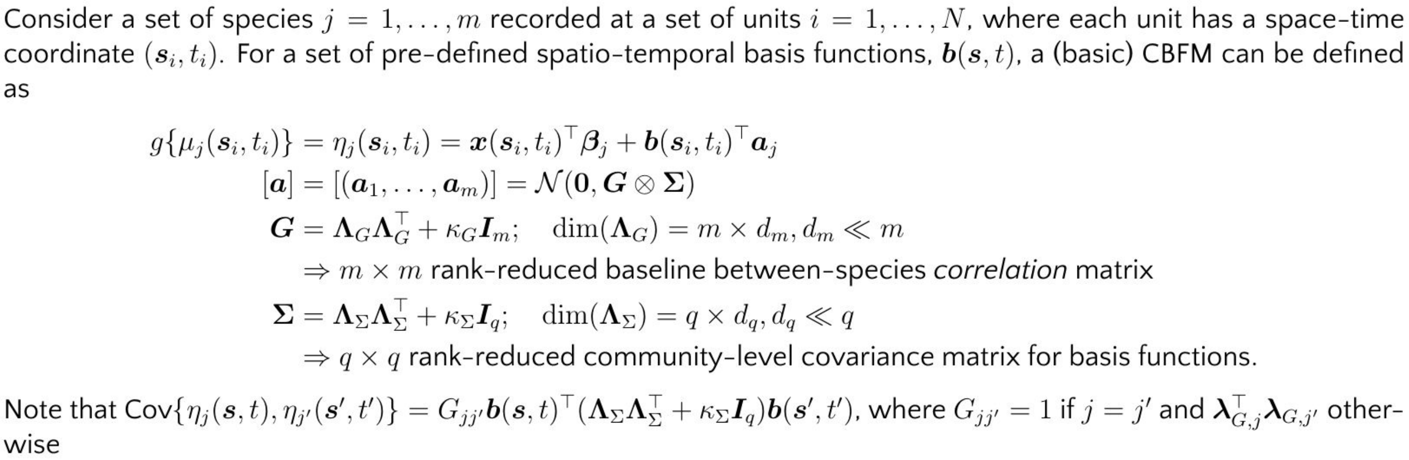
Special Issue: Time Series in the Environmental Sciences

July 2011

Pages 430-446

Citation Statements beta

Supporting	Mentioning	Contrasting
✓ 0	71	? 0



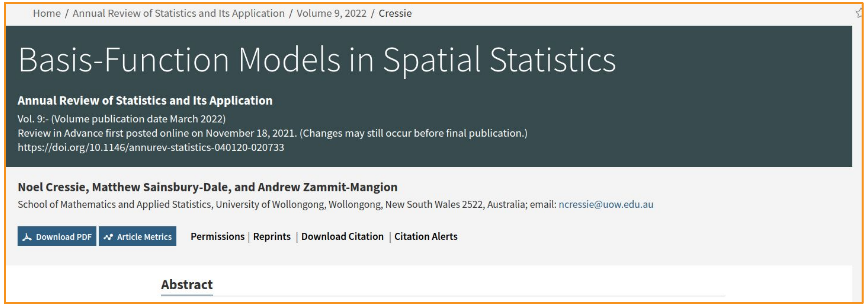
Consider a set of species $j = 1, \ldots, m$ recorded at a set of units $i = 1, \ldots, N$, where each unit has a space-time coordinate $(\boldsymbol{s}_i, t_i)$. For a set of pre-defined spatio-temporal basis functions, $\boldsymbol{b}(\boldsymbol{s}, t)$, a (basic) CBFM can be defined as

$$
\begin{aligned}
g\{\mu_j(\boldsymbol{s}_i, t_i)\} &= \eta_j(\boldsymbol{s}_i, t_i) = \boldsymbol{x}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{\beta}_j + \boldsymbol{b}(\boldsymbol{s}_i, t_i)^\top \boldsymbol{a}_j \\
[\boldsymbol{a}] &= [(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)] = \mathcal{N}(\boldsymbol{0}, \boldsymbol{G} \otimes \boldsymbol{\Sigma}) \\
\boldsymbol{G} &= \boldsymbol{\Lambda}_G \boldsymbol{\Lambda}_G^\top + \kappa_G \boldsymbol{I}_m; \quad \dim(\boldsymbol{\Lambda}_G) = m \times d_m, d_m \ll m \\
&\Rightarrow m \times m \text{ rank-reduced baseline between-species } \textit{correlation} \text{ matrix} \\
\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q; \quad \dim(\boldsymbol{\Lambda}_\Sigma) = q \times d_q, d_q \ll q \\
&\Rightarrow q \times q \text{ rank-reduced community-level covariance matrix for basis functions.}
\end{aligned}
$$

Note that Cov$\{\eta_j(\boldsymbol{s}, t), \eta_{j'}(\boldsymbol{s}', t')\} = G_{jj'} \boldsymbol{b}(\boldsymbol{s}, t)^\top (\boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top + \kappa_\Sigma \boldsymbol{I}_q) \boldsymbol{b}(\boldsymbol{s}', t')$, where $G_{jj'} = 1$ if $j = j'$ and $\boldsymbol{\lambda}_{G,j}^\top \boldsymbol{\lambda}_{G,j'}$ otherwise

Gen 3: CBFMs?

- Basis functions are not new news
  - GAMs have been known in ecology for a long time. But not so much fixed rank kriging



Home / Annual Review of Statistics and Its Application / Volume 9, 2022 / Cressie

Basis-Function Models in Spatial Statistics

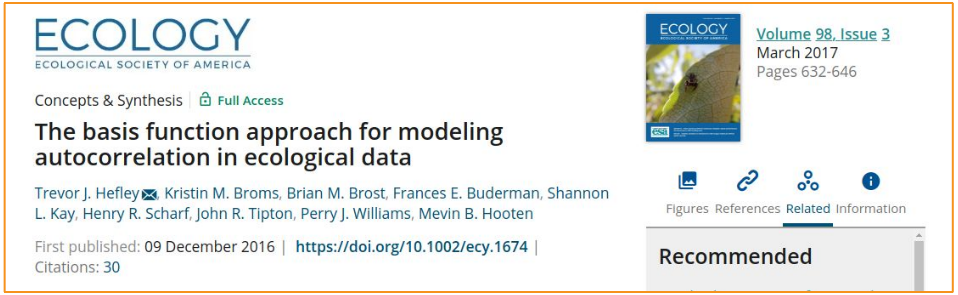
Annual Review of Statistics and Its Application

Vol. 9:- (Volume publication date March 2022)
Review in Advance first posted online on November 18, 2021. (Changes may still occur before final publication.)
https://doi.org/10.1146/annurev-statistics-040120-020733

Noel Cressie, Matthew Sainsbury-Dale, and Andrew Zammit-Mangion

School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, New South Wales 2522, Australia; email: ncressie@uow.edu.au

Download PDF
Article Metrics
Permissions | Reprints | Download Citation | Citation Alerts
Abstract



ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

Concepts & Synthesis | Full Access

The basis function approach for modeling autocorrelation in ecological data

Trevor J. Hefley, Kristin M. Broms, Brian M. Brost, Frances E. Buderman, Shannon L. Kay, Henry R. Scharf, John R. Tipton, Perry J. Williams, Mevin B. Hooten

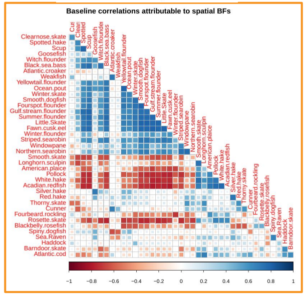
First published: 09 December 2016 | https://doi.org/10.1002/ecy.1674 |
Citations: 30

Volume 98, Issue 3
March 2017
Pages 632-646

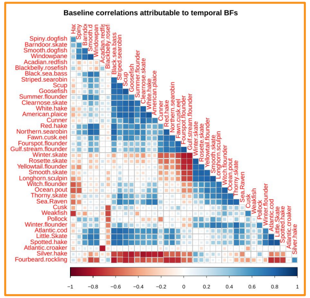
Gen 3: CBFMs?

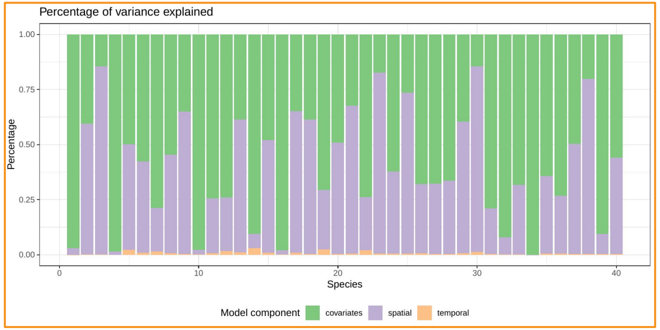
- Basis functions are not new news
  - GAMs have been known in ecology for a long time. But not so much fixed rank kriging
  - It takes a while to translate statistical methods to other disciplines (properly)...

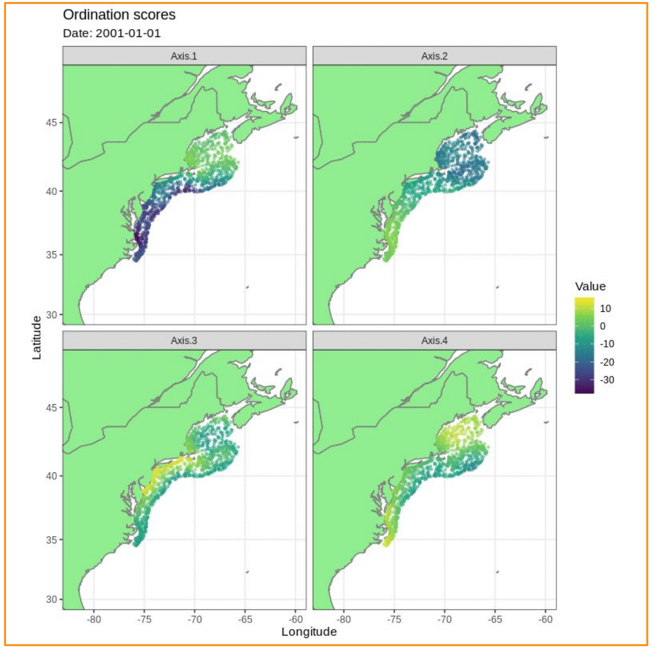
$$G_{\mathcal{S}} = \Lambda_{G_{\mathcal{S}}} \Lambda_{G_{\mathcal{S}}}^{\top} + \kappa_{G_{\mathcal{S}}} I_m$$



$$G_{\mathcal{T}} = \Lambda_{G_{\mathcal{T}}} \Lambda_{G_{\mathcal{T}}}^{\top} + \kappa_{G_{\mathcal{T}}} I_m$$







Gen 3: CBFMs?

- Basis functions are not new news
  - GAMs have been known in ecology for a long time. But not so much fixed rank kriging
  - It takes a while to translate statistical methods to other disciplines (properly)...

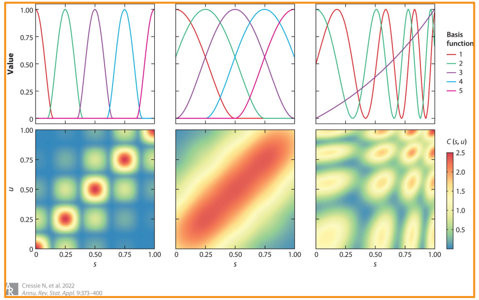
- Is Gen 3 ⊂ Gen2+? Isn't basis functions just an approximation of spatio-temporal LVMs?

"Sales are up 38% since the name change."

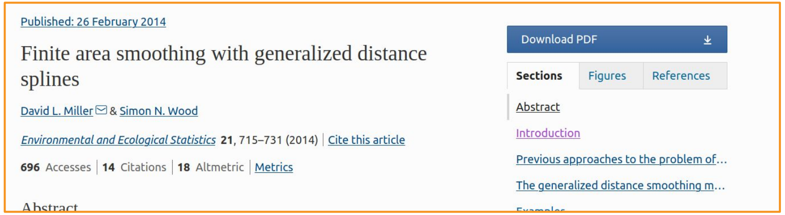
Gen 3: CBFMs?

- Basis functions are not new news
  - GAMs have been known in ecology for a long time. But not so much fixed rank kriging
  - It takes a while to translate statistical methods to other disciplines (properly)…

- Is Gen 3 ⊂ Gen2+? Isn't basis functions just an approximation of spatio-temporal LVMs?
  - Depends on how you want to approach basis functions: *"one person's mean is another person's covariance"* (Cressie, 1993)
  - A "basis function" mindset can opens up new opportunities



Cressie N, et al. 2022
Annu. Rev. Stat. Appl. 9:373–400



Published: 26 February 2014

Finite area smoothing with generalized distance splines

David L. Miller & Simon N. Wood

Environmental and Ecological Statistics **21**, 715–731 (2014) | Cite this article

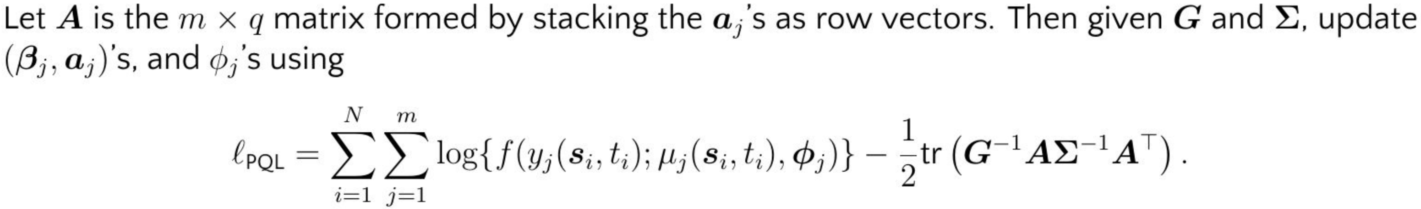
696 Accesses | **14** Citations | **18** Altmetric | Metrics

Abstract

Abstract
Introduction
Previous approaches to the problem of…
The generalized distance smoothing m…
Examples

Estimation, inference and all that jazz

- CBFM = Leveled up FRK = A very big generalized additive model (GAM)
  - Penalized quasi-likelihood (PQL) estimation for all coefficients, dispersion parameters; amenable to parallelization

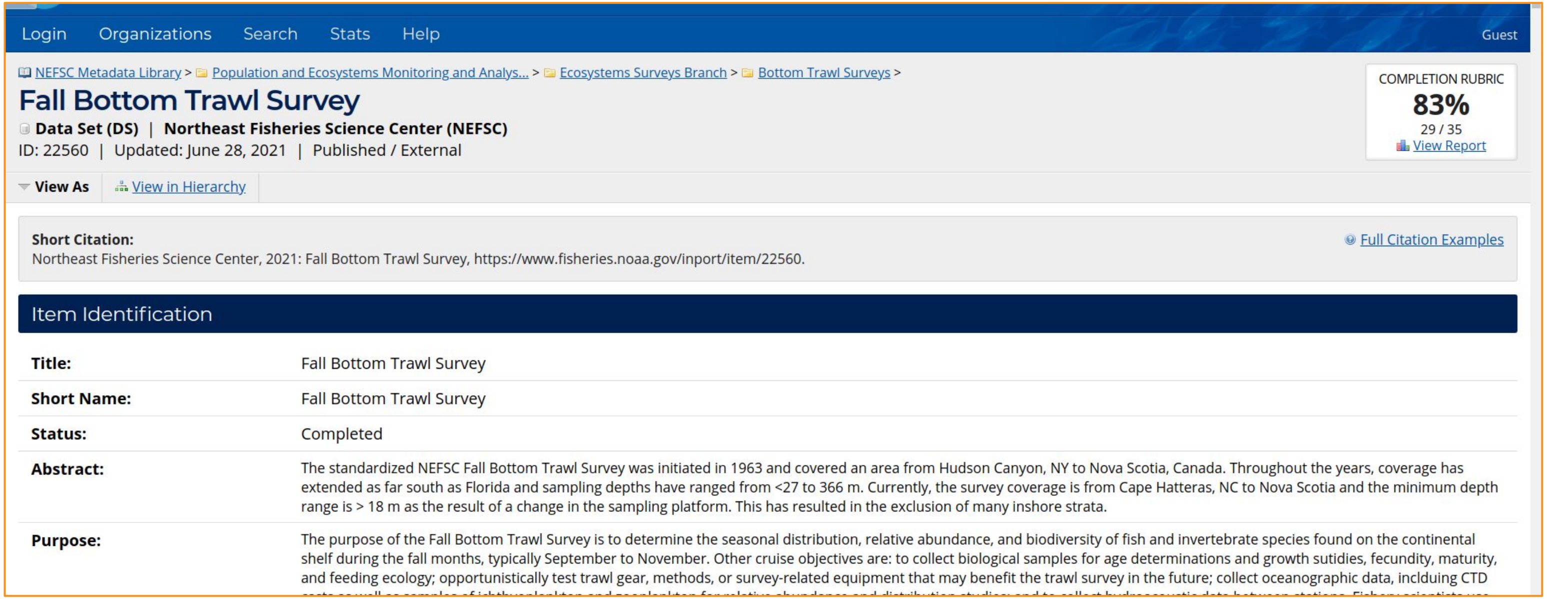


Let $\boldsymbol{A}$ is the $m \times q$ matrix formed by stacking the $\boldsymbol{a}_j$'s as row vectors. Then given $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$, update $(\boldsymbol{\beta}_j, \boldsymbol{a}_j)$'s, and ϕ_j's using

$$\ell_{\mathrm{PQL}} = \sum_{i=1}^{N} \sum_{j=1}^{m} \log\{f(y_j(\boldsymbol{s}_i, t_i); \mu_j(\boldsymbol{s}_i, t_i), \boldsymbol{\phi}_j)\} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{G}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^{\top}\right).$$

Multivariate abundance data

- NorthEast Fisheries Science Center (NEFSC) fall bottom trawl survey
  - https://www.fisheries.noaa.gov/inport/item/22560
  - Subset of 2000-2019



NEFSC Metadata Library > Population and Ecosystems Monitoring and Analys... > Ecosystems Surveys Branch > Bottom Trawl Surveys >

Fall Bottom Trawl Survey

Data Set (DS) | **Northeast Fisheries Science Center (NEFSC)**
ID: 22560 | Updated: June 28, 2021 | Published / External

COMPLETION RUBRIC
83%
29 / 35
View Report

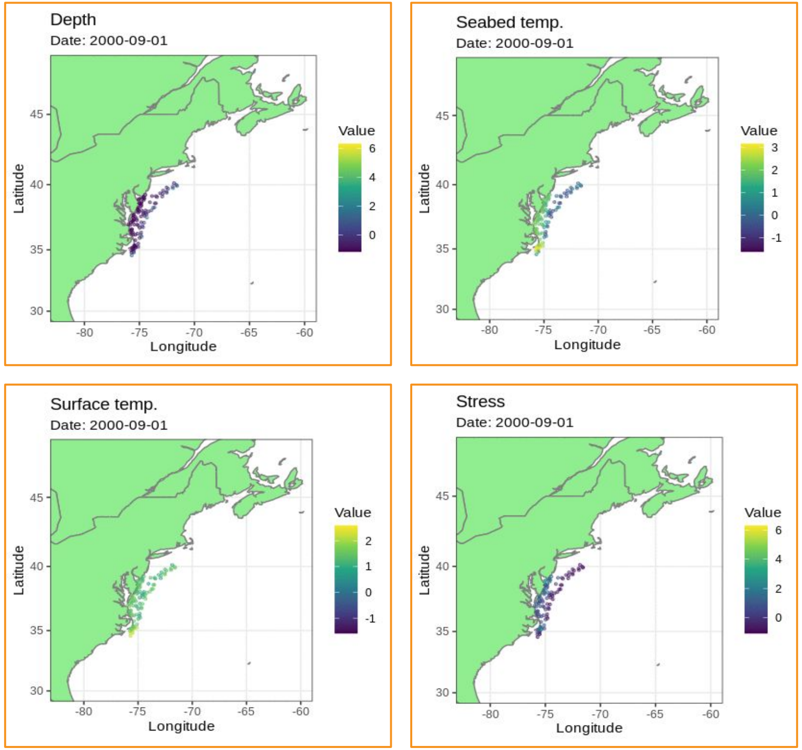
Short Citation:
Northeast Fisheries Science Center, 2021: Fall Bottom Trawl Survey, https://www.fisheries.noaa.gov/inport/item/22560.

Item Identification

Title:	Fall Bottom Trawl Survey
Short Name:	Fall Bottom Trawl Survey
Status:	Completed
Abstract:	The standardized NEFSC Fall Bottom Trawl Survey was initiated in 1963 and covered an area from Hudson Canyon, NY to Nova Scotia, Canada. Throughout the years, coverage has extended as far south as Florida and sampling depths have ranged from <27 to 366 m. Currently, the survey coverage is from Cape Hatteras, NC to Nova Scotia and the minimum depth range is > 18 m as the result of a change in the sampling platform. This has resulted in the exclusion of many inshore strata.
Purpose:	The purpose of the Fall Bottom Trawl Survey is to determine the seasonal distribution, relative abundance, and biodiversity of fish and invertebrate species found on the continental shelf during the fall months, typically September to November. Other cruise objectives are: to collect biological samples for age determinations and growth sutidies, fecundity, maturity, and feeding ecology; opportunistically test trawl gear, methods, or survey-related equipment that may benefit the trawl survey in the future; collect oceanographic data, inclduing CTD

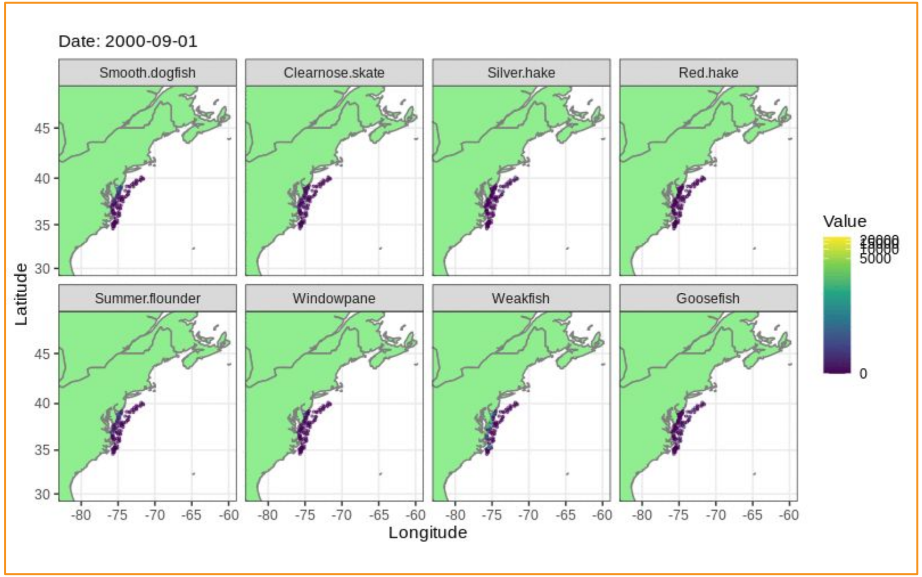
Multivariate abundance data

- Four example covariates in fall bottom trawl survey:
    - There are more covariates (between 20–30)…



Multivariate abundance data

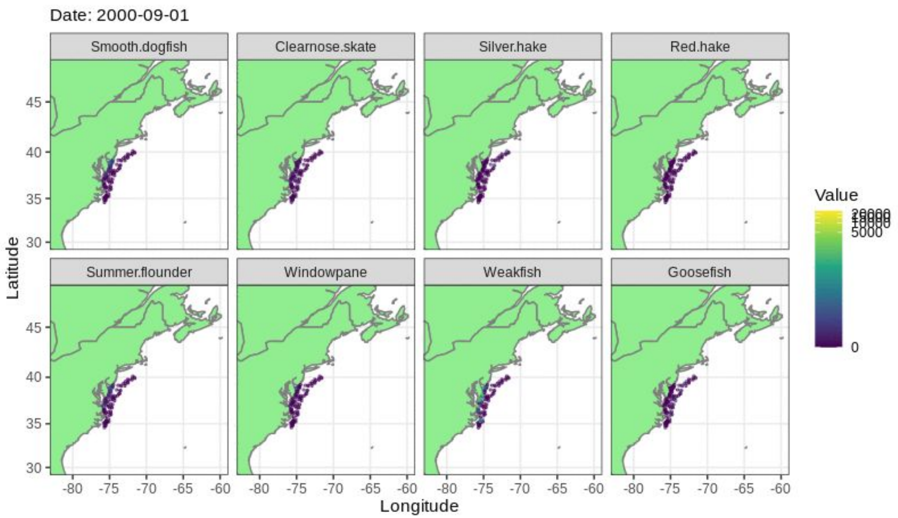
- Eight example demersal fish species in fall bottom trawl survey
    - Around 150ish taxa in total
    - High-dimensional, correlated responses



Multivariate abundance data

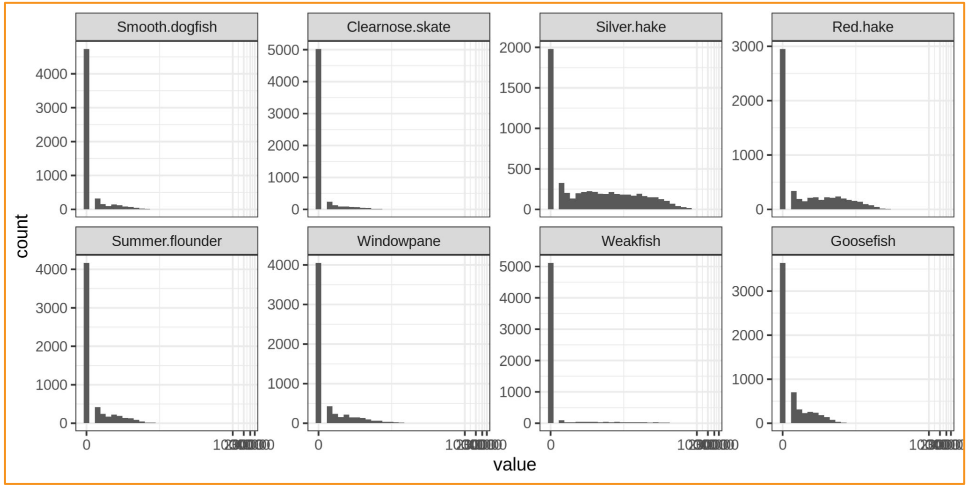
- Eight example demersal fish species in fall bottom trawl survey
    - Around 150ish taxa in total
    - High-dimensional, correlated responses

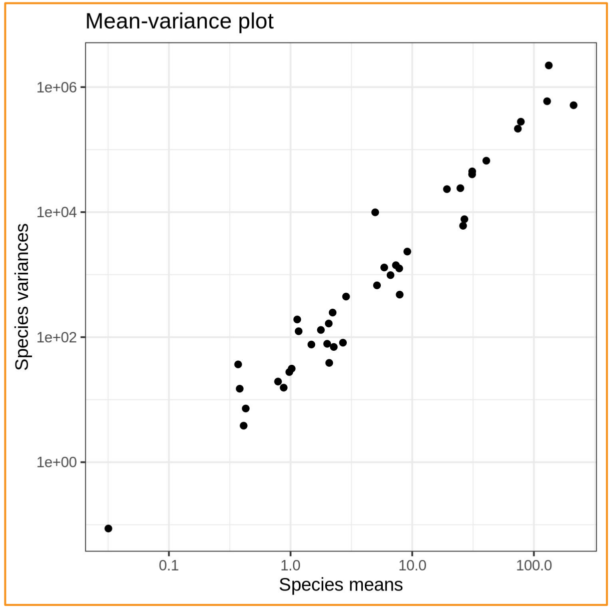
- Some other noteworthy points:
    - You never visit the same location more than once
    - About 6,000 space-time locations visited between 2000-2019



Multivariate abundance data

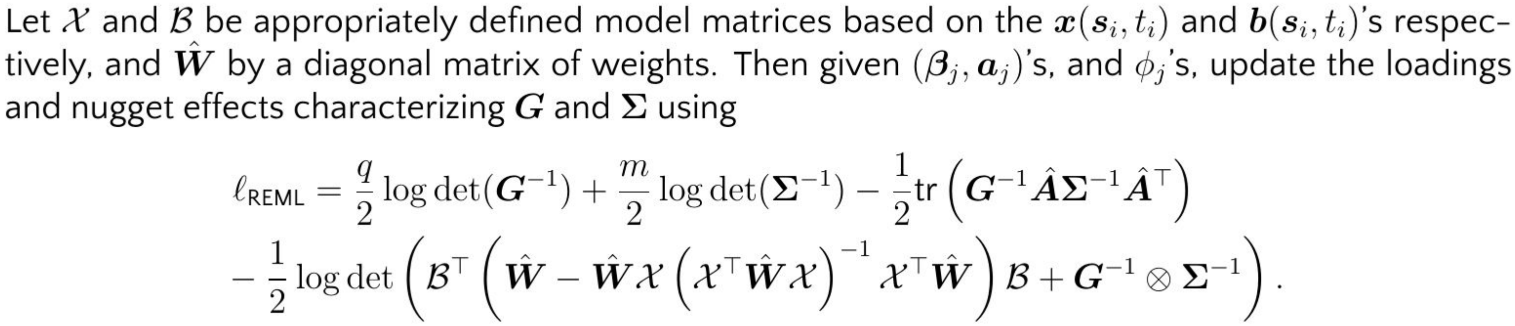
- Responses are:
  - Sparse, non-continuous
  - Strong mean-variance relationship (various reasons behind this)





Estimation, inference and all that jazz

- CBFM = Leveled up FRK = A very big generalized additive model (GAM)
  - Penalized quasi-likelihood (PQL) estimation for all coefficients, dispersion parameters; amenable to parallelization
  - Maximum restricted Laplace-approximated likelihood estimation for the loadings and nugget effects

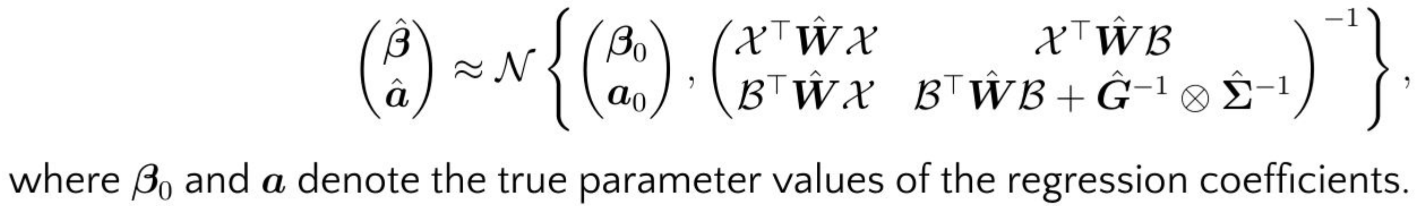


Let $\mathcal{X}$ and $\mathcal{B}$ be appropriately defined model matrices based on the $\boldsymbol{x}(\boldsymbol{s}_i, t_i)$ and $\boldsymbol{b}(\boldsymbol{s}_i, t_i)$'s respectively, and $\hat{\boldsymbol{W}}$ by a diagonal matrix of weights. Then given $(\boldsymbol{\beta}_j, \boldsymbol{a}_j)$'s, and ϕ_j's, update the loadings and nugget effects characterizing $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ using

$$\ell_{\text{REML}} = \frac{q}{2}\log\det(\boldsymbol{G}^{-1}) + \frac{m}{2}\log\det(\boldsymbol{\Sigma}^{-1}) - \frac{1}{2}\text{tr}\left(\boldsymbol{G}^{-1}\hat{\boldsymbol{A}}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{A}}^\top\right)$$
$$- \frac{1}{2}\log\det\left(\mathcal{B}^\top\left(\hat{\boldsymbol{W}} - \hat{\boldsymbol{W}}\mathcal{X}\left(\mathcal{X}^\top\hat{\boldsymbol{W}}\mathcal{X}\right)^{-1}\mathcal{X}^\top\hat{\boldsymbol{W}}\right)\mathcal{B} + \boldsymbol{G}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right).$$

Estimation, inference and all that jazz

- CBFM = Leveled up FRK = A very big generalized additive model (GAM)
  - Penalized quasi-likelihood (PQL) estimation for all coefficients, dispersion parameters; amenable to parallelization
  - Maximum restricted Laplace-approximated likelihood estimation for the loadings and nugget effects
  - Approximate large sample distributions for coefficients/linear predictors etc...

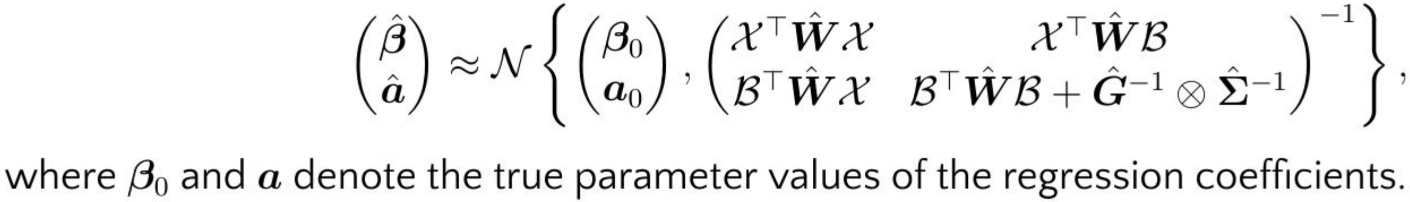


$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{a}}\end{pmatrix} \approx \mathcal{N}\left\{\begin{pmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{a}_0\end{pmatrix}, \begin{pmatrix}\mathcal{X}^\top\hat{\mathbf{W}}\mathcal{X} & \mathcal{X}^\top\hat{\mathbf{W}}\mathcal{B}\\ \mathcal{B}^\top\hat{\mathbf{W}}\mathcal{X} & \mathcal{B}^\top\hat{\mathbf{W}}\mathcal{B}+\hat{\boldsymbol{G}}^{-1}\otimes\hat{\boldsymbol{\Sigma}}^{-1}\end{pmatrix}^{-1}\right\},$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{a}$ denote the true parameter values of the regression coefficients.

Estimation, inference and all that jazz

- CBFM = Leveled up FRK = A very big generalized additive model (GAM)
  - Penalized quasi-likelihood (PQL) estimation for all coefficients, dispersion parameters; amenable to parallelization
  - Maximum restricted Laplace-approximated likelihood estimation for the loadings and nugget effects
  - Approximate large sample distributions for coefficients/linear predictors etc...
  - Adapt GAM tools for residual analysis, model selection, prediction etc...; variance-partitioning; space-time ordination using SVD-type ideas, and so on



$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{a}}\end{pmatrix} \approx \mathcal{N}\left\{\begin{pmatrix}\boldsymbol{\beta}_0\\ \boldsymbol{a}_0\end{pmatrix}, \begin{pmatrix}\mathcal{X}^\top\hat{\boldsymbol{W}}\mathcal{X} & \mathcal{X}^\top\hat{\boldsymbol{W}}\mathcal{B}\\ \mathcal{B}^\top\hat{\boldsymbol{W}}\mathcal{X} & \mathcal{B}^\top\hat{\boldsymbol{W}}\mathcal{B}+\hat{\boldsymbol{G}}^{-1}\otimes\hat{\boldsymbol{\Sigma}}^{-1}\end{pmatrix}^{-1}\right\},$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{a}$ denote the true parameter values of the regression coefficients.