Small area estimation of zero-inflated, spatially correlated forest variables using copula models

2019 FIA Stakeholders Science Meeting

Vicente Monleon, Lisa Madsen and Lisa Wilson

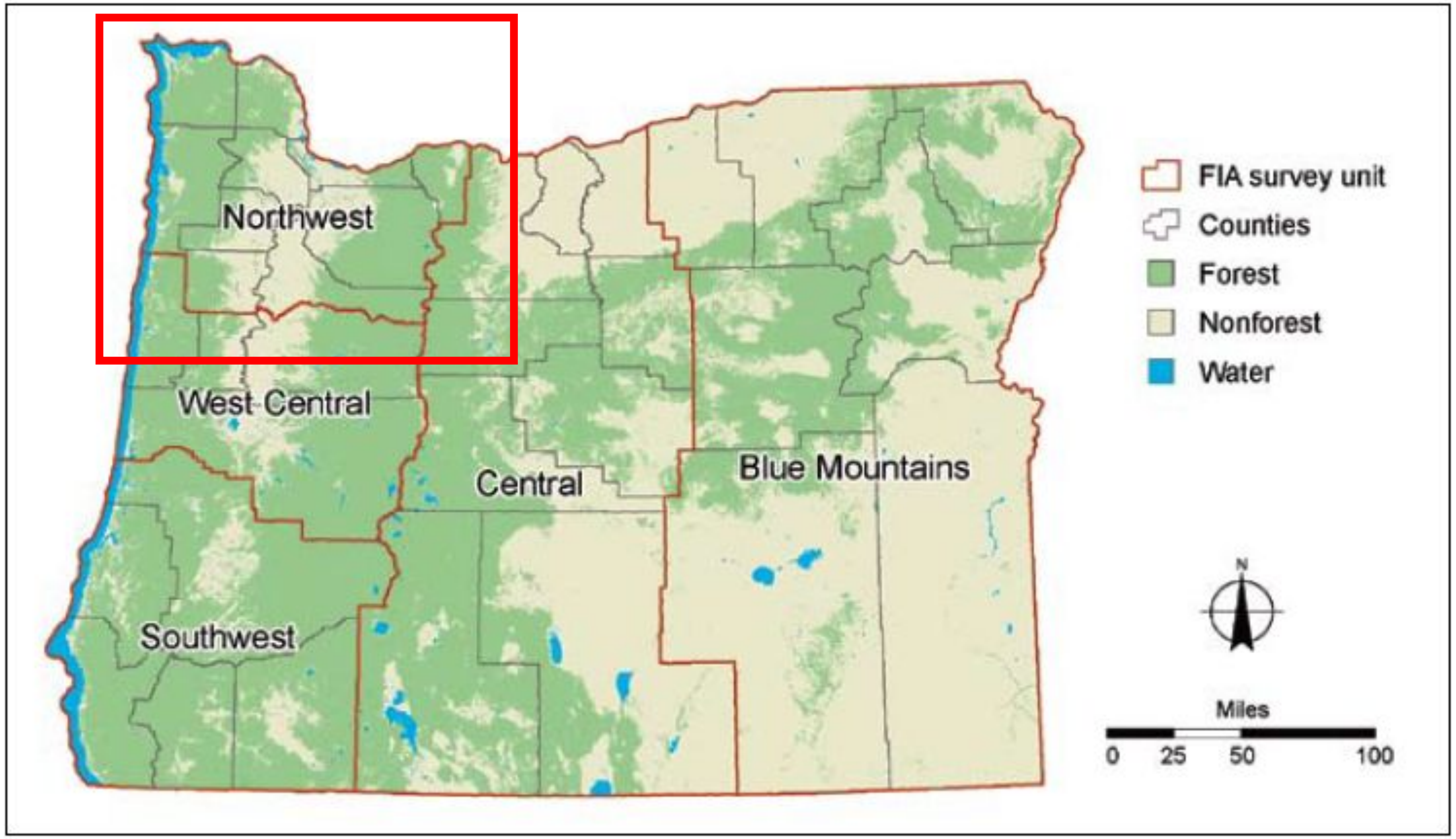
USDA Forest Service, PNW Research Station

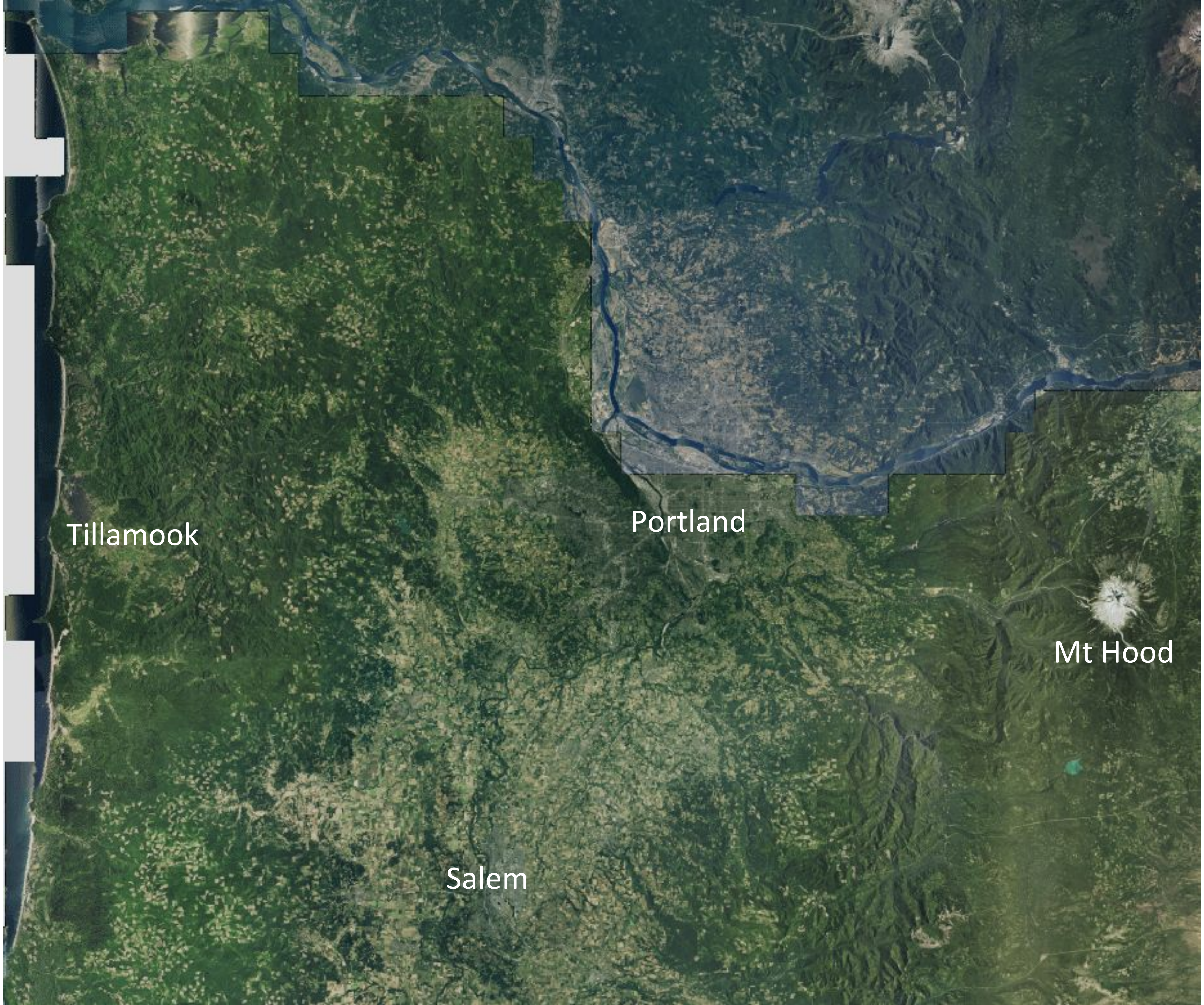
Department of Statistics, Oregon State University

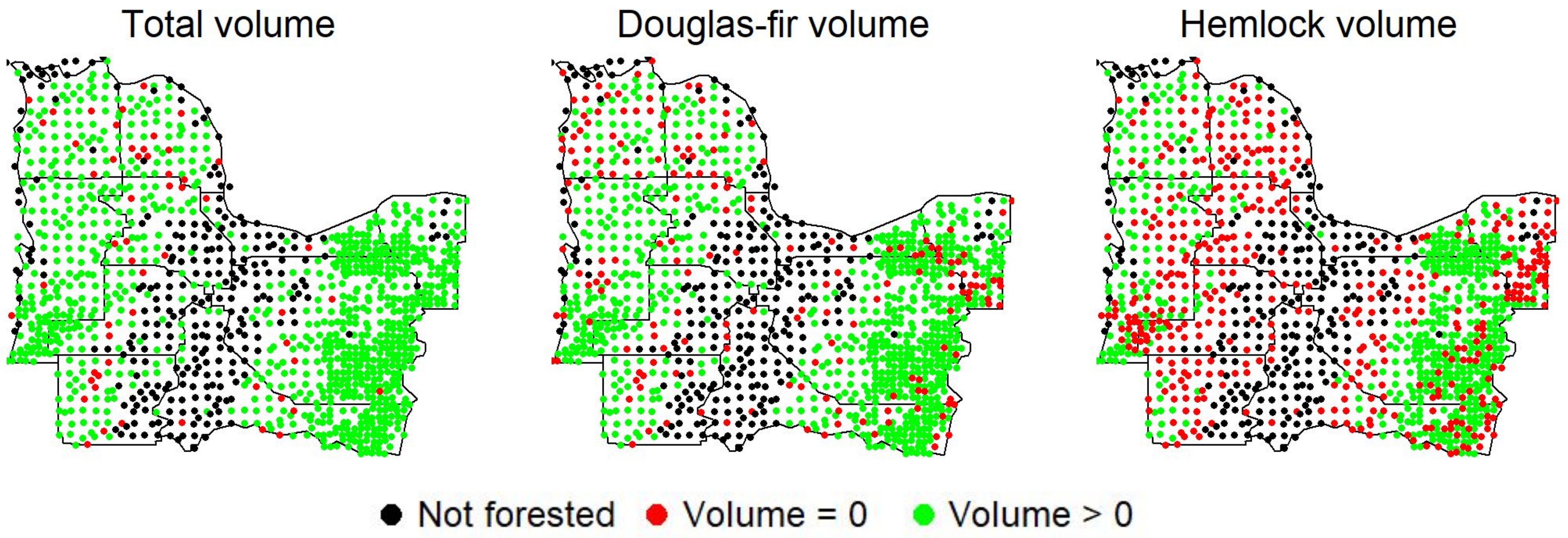
Modelling forest inventory variables is challenging

- Spatial correlation: locations close to each other “share” information
  - each plot does not represent a “full” unit of information.
  - the “shared” information can be used to improve prediction (i.e., kriging)
- Zero-inflated: a large proportion of the values of the variable are 0
  - Non-forest land, harvested areas, species not present…
  - The proportion of 0s increases as the domain becomes more restricted
- Often positive, very skewed
- This precludes using traditional modelling approaches based on the normal distribution

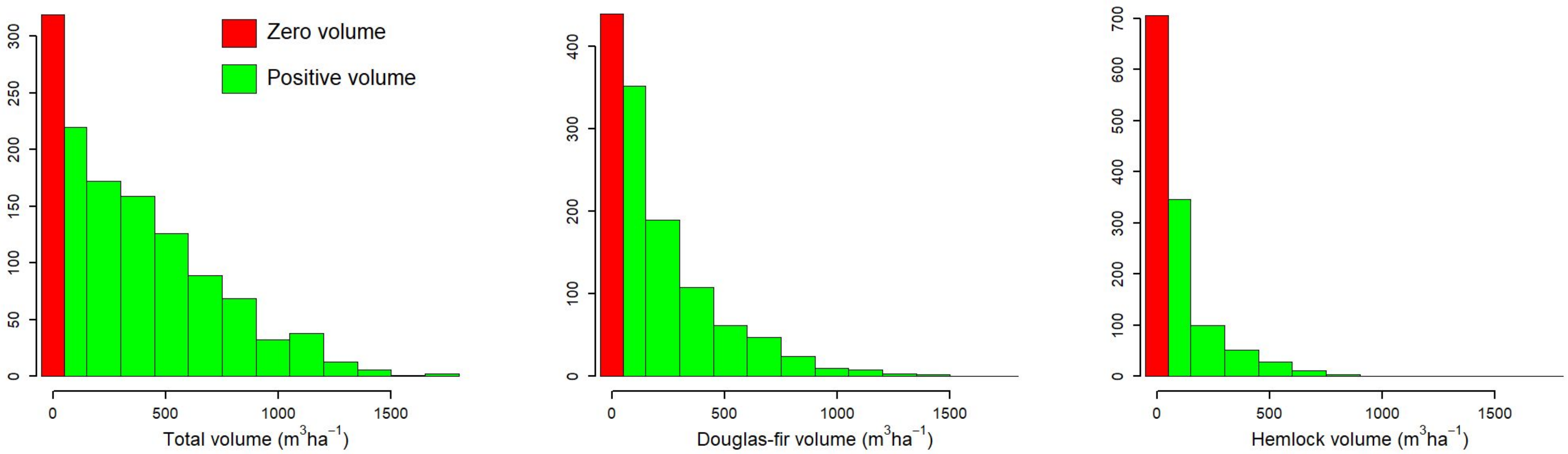
Example: timber volume in NW Oregon




Tillamook
Portland
Mt Hood
Salem



	Total	Douglas-fir	Hemlock
Percentage of plots with 0 volume	26	35	57

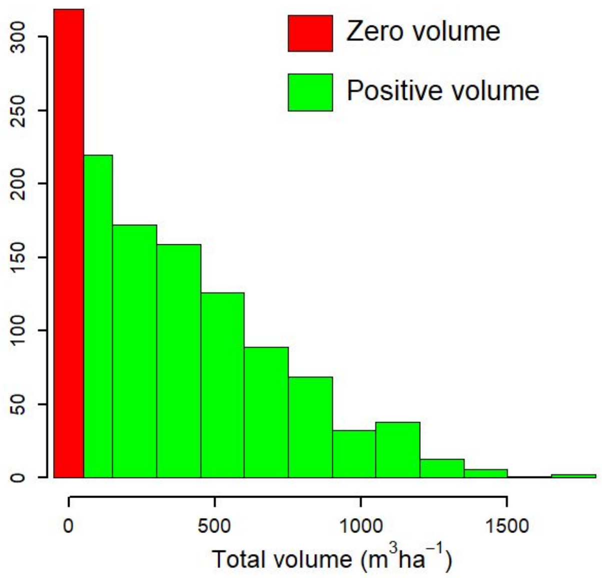


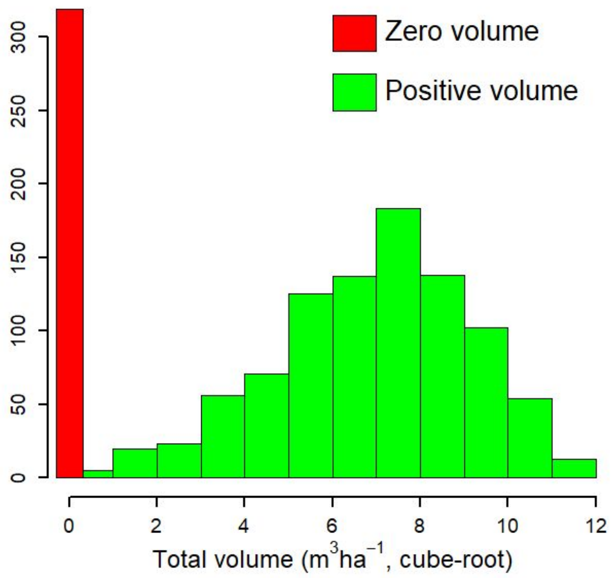
- Not only it is zero-inflated, but the distribution is highly skewed
- No simple transformation / solution can deal with those problems

What is a copula model?

- A copula is a multivariate distribution function for which the marginal probability distribution of each variable is uniform.
- Take advantage of the probability integral transformation
  - If V is a random variable with cumulative distribution function $F_V(v)$, then the variable $U = F_V(v)$ is uniformly distributed on $(0,1)$ [i.e., $P(U \leq u) = u$]
- The marginal distributions and the copula can be examined separately and fitted either separately or jointly using maximum likelihood.
- Main result (Sklar): every multivariate distribution function can be expressed in terms of its univariate marginal distributions and a copula describing the dependence among them.

Marginal distribution





- A cubic root transformation of the non-zero volumes worked best

Marginal distribution: zero-inflated gamma

- Zero-inflated gamma model to account for the excess 0s
- The observed volume (V, cube root) is a Bernoulli mixture of a 0 and a Gamma random variable:

$B \sim \mathrm{Bernoulli}(\pi)$ $\quad$ π: probability of volume > 0

$W \sim \mathrm{gamma}(\alpha, \beta)$ $\quad$ W: volume (cube root), given that it is not 0

$$V = (1 - B) \cdot 0 + B \cdot W$$

- Modelled the mean of B and W as a function of an indicator of forestland (based on of NLCD forest cover classes), Landsat tesseled cap "wetness" variable (tsc3), and their interaction

Gaussian copula: double transformation

- First transformation: estimate the cumulative distribution function of this marginal distribution, $U = F_V(v)$
- Second transformation: univariate standard normal, $\Phi^{-1}\left(F_V(v)\right)$ [Φ is the standard normal cumulative distribution function]
- Join together in a multivariate standard normal distribution
- Model the spatial dependence structure via a (rank) correlation matrix

$$C(v; \Sigma) = \Phi_\Sigma\left[\Phi^{-1}\left(F_1(v_1)\right), \dots, \Phi^{-1}\left(F_n(v_n)\right)\right]$$

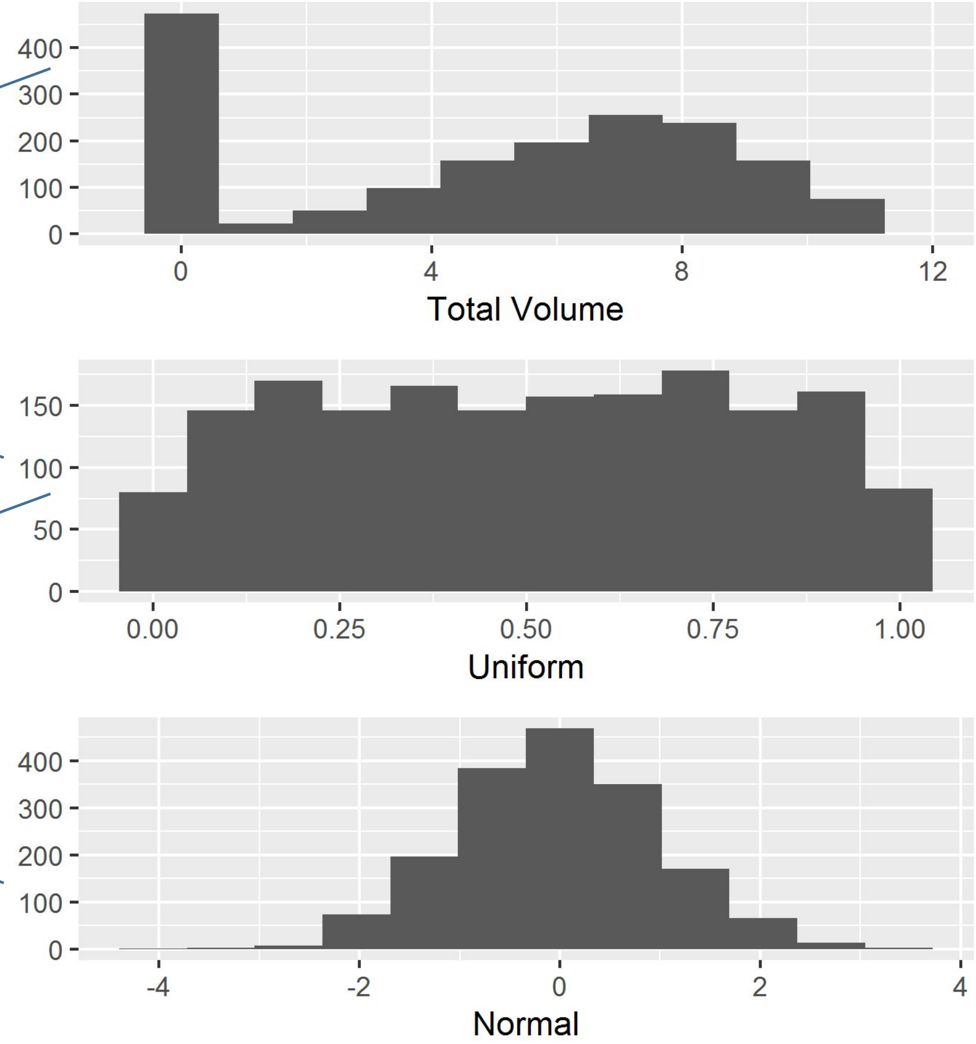
[Φ_Σ multivariate normal with mean 0 and correlation matrix Σ]

- Use the normal distribution tools for analysis /prediction (kriging)
- Reverse the transformations to the original scale

Transformation to Normal

$U = \mathrm{cdf}(V)$

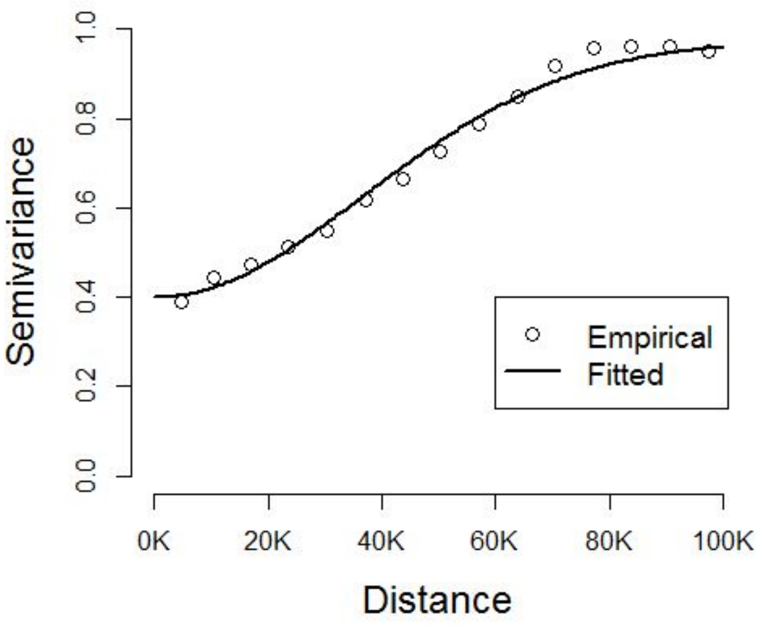
$Z = \text{inverse normal cdf}(U)$



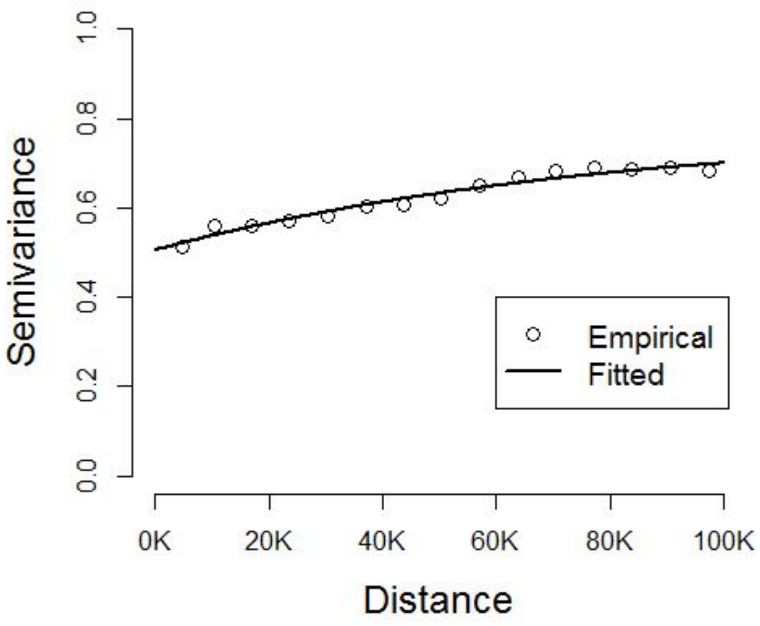
These steps are reversible.

Results

- For total volume, once we include the covariates, the spatial correlation becomes negligible



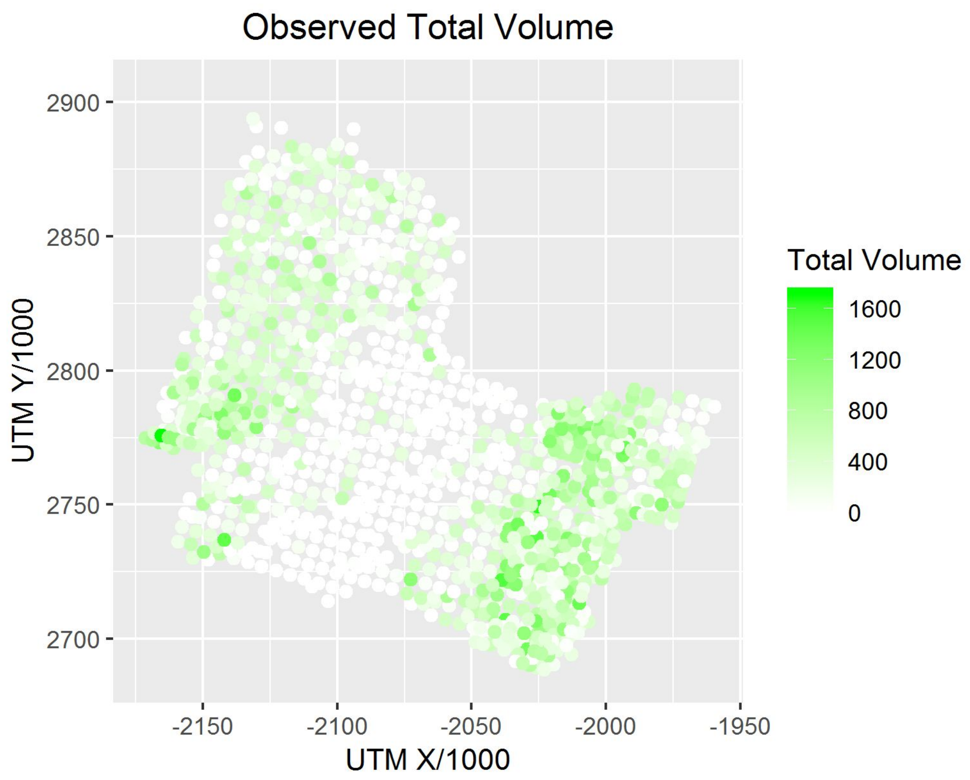
No-covariate model

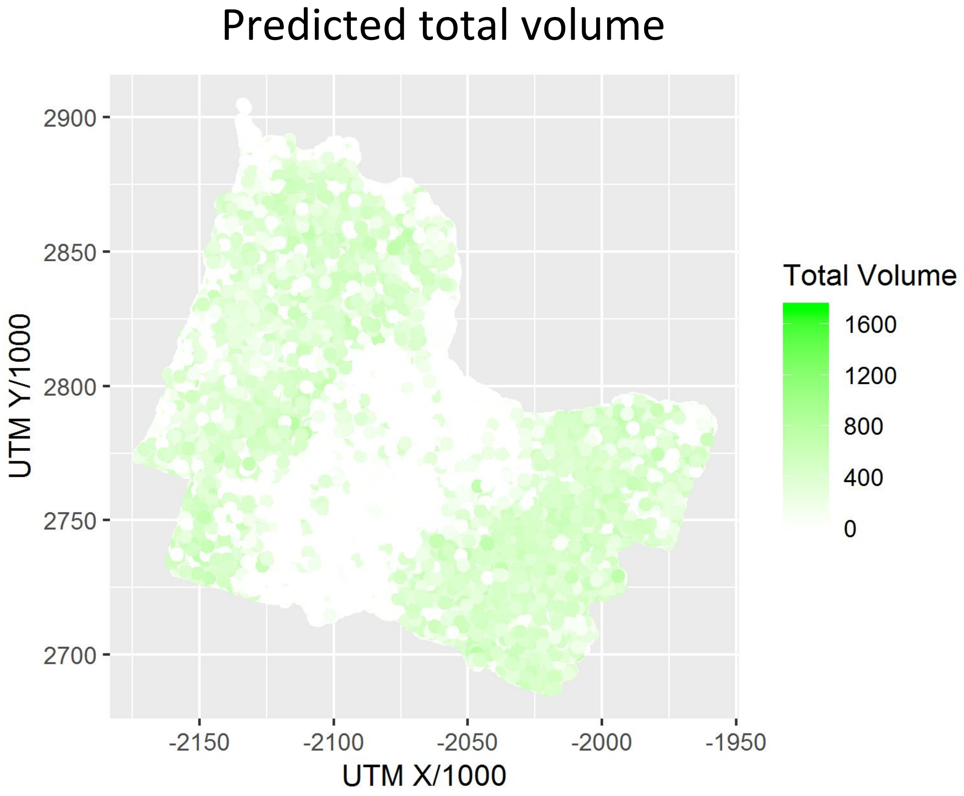


Modelled data

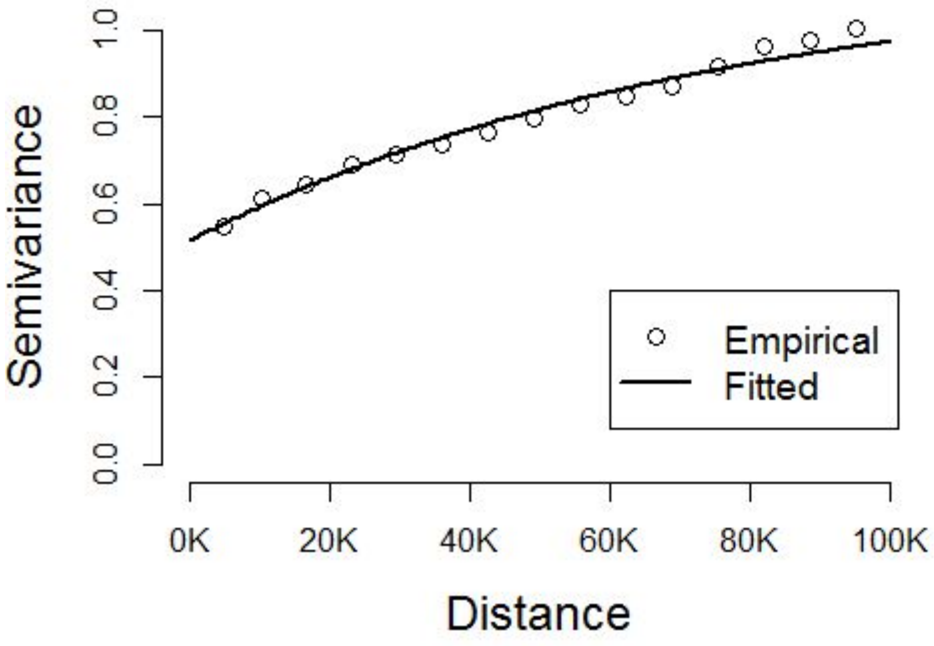
Total volume predictions

- Since spatial correlation is very weak, we can use a simple zero-inflated gamma model

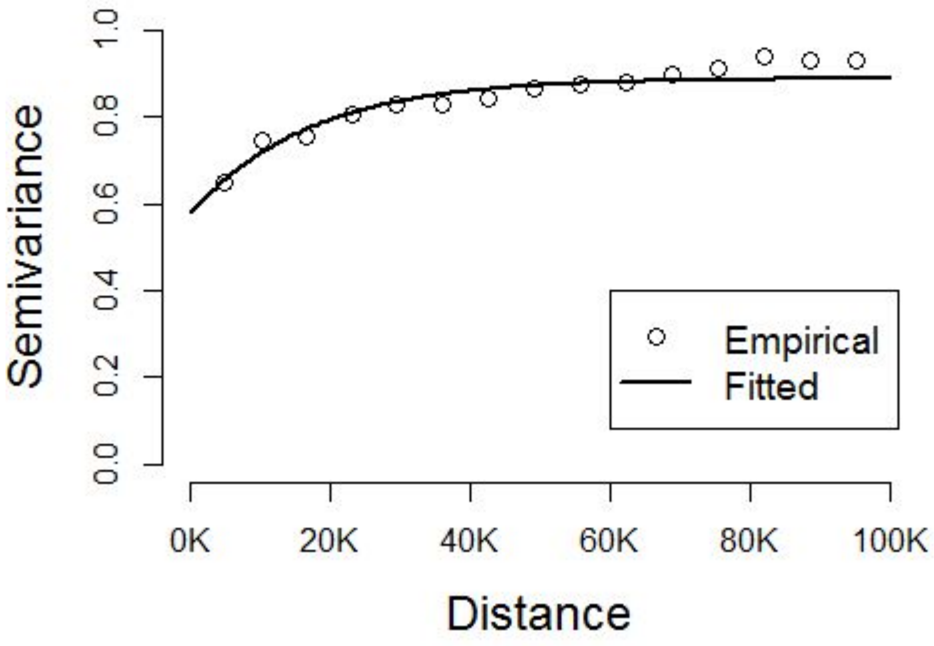




Hemlock volume - semivariograms

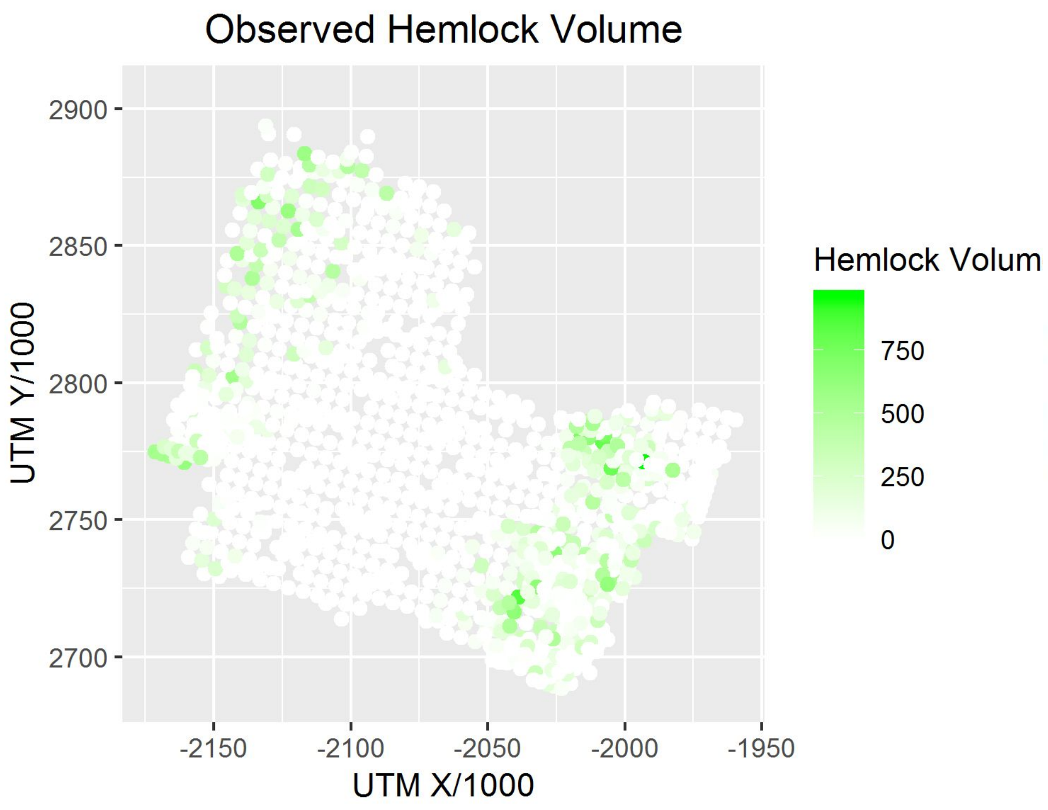


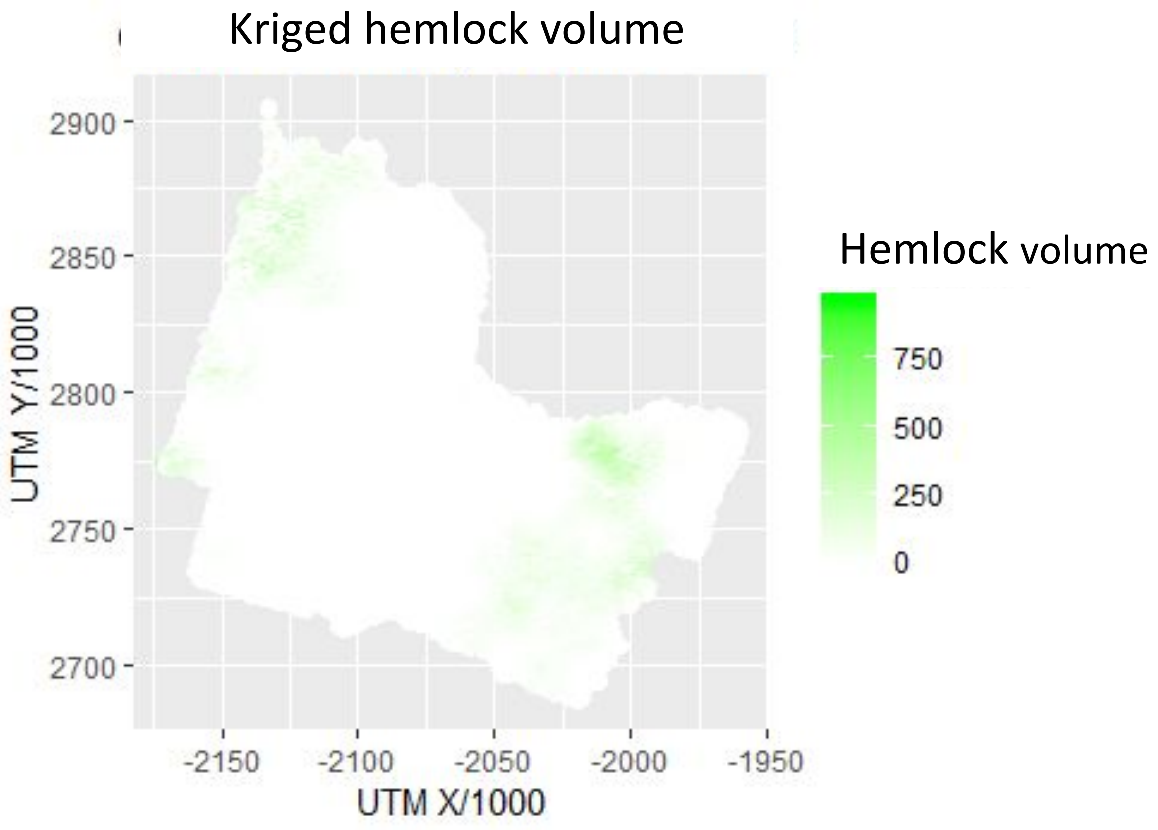
No-covariate model



Modelled data

Hemlock volume predictions





Conclusions and future work

- Gaussian copulas allow us to build realistic models for forest inventory variables, incorporating spatial correlation and non-standard distributions.
- Add covariates to build operational models
- Small area estimation: how to compute measures of uncertainty in the original scale
- High dimensional dataset, computational difficulties.