Statistical upscaling of terrestrial greenhouse gas emissions

Gerard Heuvelink, Wim De Vries, Tom Hoogland, Hans Kros and Gert Jan Reinds
Upscaling

- Upscaling is taken as synonymous to \textit{aggregation}, where the interest is in obtaining an integrated value of a variable over an area or volume of a given size and/or a time interval of given length.

- Example: chambers may measure emission over 30 minutes for a 0.25 m$^2$ surface, whereas the interest is in annual values for a large region or country.

- Upscaling is based on ‘point’ observations, but perhaps auxiliary information can be used to improve accuracy.
Two fundamentally different approaches to upscaling

1. **Design-based**
   - Makes no assumption about space-time variability structure (‘model-free’)
   - Does not suffer from making wrong assumptions
   - Locations of observations must be selected with probability sampling, usually simple random sampling
   - Simple random sampling may be replaced with more efficient designs (e.g. stratified sampling, two-phase random sampling, cluster sampling)
   - Widespread misconception that design-based methods cannot be applied when there is spatial (temporal) correlation
   - Problems with scarce data and preferential sampling
Two fundamentally different approaches to upscaling

1. Model-based

- Assumes a (statistical) model that characterises the space-time behaviour of the variable of interest
- **Statistical** model because space-time behaviour is partially unpredictable: include stochastic term to represent uncertainty
- Model also includes a deterministic trend, ranging from an unknown constant to a complex process model such as DNDC: $Z(x,t) = m(x,t) + \varepsilon(x,t)$
- Given the model, trend and observations, estimates of upscaled variable are obtained with block-kriging
- Model-based more flexible and potentially more efficient than design-based, but makes assumptions
Model-based upscaling: two main steps

1. point-support data at measurement locations
2. spatial coverage of point-support data
   - interpretation
   - aggregation
3. spatial coverage of block-support data
Example design-based upscaling

- Aggregation over time only: from half-hour chamber measurements to annual average (1 July 2001 to 30 June 2002)
- $\text{N}_2\text{O}$ measured at 26 times for two grassland parcels (dry and wet) in Western Dutch peat soil area
- Assume stratified random sampling with two strata: growing season (1 March to 30 September) and non-growing season
- Higher sampling density in growing season
Measurements over time

Emission in microgr N/m²/hour

Day number (1 July 2001 = 1)
Box plots show differences between plots and season.
Statistical inference

\[ \mu = \frac{T_g}{T} \mu_g + \frac{T_n}{T} \mu_n \]

\[ \hat{\sigma}_g^2 = s_g^2 = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (x_{gi} - \bar{X}_g)^2 \]

\[ \hat{\mu} = \frac{T_g}{T} \hat{\mu}_g + \frac{T_n}{T} \hat{\mu}_n \]

\[ \sigma(\mu_g - \hat{\mu}_g) = \frac{\sigma_g}{\sqrt{n_g}} \]

\[ \sigma(\mu - \hat{\mu}) = \sqrt{\left( \frac{T_g}{T} \right)^2 \frac{\sigma_g^2}{n_g} + \left( \frac{T_n}{T} \right)^2 \frac{\sigma_n^2}{n_n}} \]
Results (recall n=26)

\[ \hat{\mu}_{\text{dry}} = 168.6 \]
\[ \sigma(\mu_{\text{dry}} - \hat{\mu}_{\text{dry}}) = 31.8 \]

\[ \hat{\mu}_{\text{wet}} = 112.5 \]
\[ \sigma(\mu_{\text{wet}} - \hat{\mu}_{\text{wet}}) = 40.6 \]
Example model-based upscaling

- $\text{N}_2\text{O}$ emission in natural areas in Europe, data from (after screening 115 observations remain):
  - NOFRETETE
  - Stehfest and Bouwman
  - Denier van der Gon

- Candidate predictors (auxiliary information incorporated in trend):
  - Climate (precipitation, nr frost days, temperature)
  - Soil (pH, organic carbon, texture, CN ratio, bulk density)
  - N deposition
  - Vegetation type (coniferous, deciduous, grass&heath)

- Use regression kriging
Regression-kriging

target variable = f(explanatory variables) + stochastic residual

Example:

\[
\log(N_2O) = \beta_0 + \beta_1 \cdot f(Prec) + \beta_2 \cdot \log(N_{dep}) + \\
\beta_3 \cdot \text{pH}_{soil} + \beta_4 \cdot \text{OrgC}_{soil} + \beta_5 \cdot \text{vegtype} + \beta_6 \cdot \text{nr}_{frost\ days} + \varepsilon
\]
Steps in regression kriging upscaling

- Build and fit regression model using emission and predictor data (auxiliary information)
- Run regression model for the whole of (natural) Europe
- Compute regression residuals at measurement locations
- Estimate spatial correlation structure of residuals
- Interpolate residuals using kriging
- Add interpolated residual to regression model output
- Slightly better approach is to integrate estimation of regression coefficients and kriging of residuals (WLS instead of OLS)
- Aggregate resulting map to desired support (e.g. compute average emission over regions or nations)
Regression model and parameter estimates based on 115 observations across Europe

\[
\log(N_2O) = \beta_0 + \beta_1 \cdot \text{Sigmoid(Prec)} + \beta_2 \cdot \log(N_{\text{dep}}) + \beta_3 \cdot \text{pH}_{\text{soil}} + \\
+ \beta_4 \cdot \text{OrgC}_{\text{soil}} + \beta_5 \cdot \text{Indicator(deciduous)} + \beta_6 \cdot \frac{\text{nr frost days}}{365} + \varepsilon
\]

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\beta_5$</th>
<th>$\beta_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.170</td>
<td>−0.246</td>
<td>0.214</td>
<td>−0.082</td>
<td>0.010</td>
<td>0.287</td>
<td>0.725</td>
</tr>
</tbody>
</table>
Predictor maps
Regression explains little variation ($R^2=0.20$)
Residual weakly spatially correlated

Residue N2O-emission [kg/ha]

- < -7
- -7 - -5
- -5 - -3
- -3 - -1
- -1 - 1
- 1 - 3
- 3 - 5
- 5 - 7
- > 7

[Map of Europe with points indicating residue N2O-emission levels.]
Regression kriging result: median $\text{N}_2\text{O}$ emission (natural areas only)

Median N2O-emission [kg/ha]
- 0 - 0.5
- 0.5 - 1
- 1 - 1.5
- 1.5 - 2
- 2 - 3
- > 3
- No Data
Large uncertainties (but note: point support!)

10 Percentile

N2O-emission [kg/ha]
- 0 - 0.25
- 0.25 - 0.5
- 0.5 - 0.75
- 0.75 - 1
- 1 - 1.5
- > 1.5
- No Data

90 Percentile
Conclusions

- Design-based upscaling attractive because it does not suffer from making wrong assumptions
- It is also suitable for validation because independence guaranteed (provided data are not used twice)
- However, measurements must be selected using probability sampling, this is rare in GHG emission research
- Time to critically evaluate measurement strategies?
- Model-based upscaling currently more suitable for GHG emission research, but model-building and data selection requires attention
- Do not expect good results with scarce and/or poor data!
Thank you