## Methodensection Hanna

## Cluster Analysis

With the reduced set of n = 209 publications, we conducted the semantic full-text cluster analysis (Abson *et al* 2014) which groups publications into different clusters based on co-abundancy of words. The rationale of this full-text cluster analysis is that publications addressing a topic in similar ways would use similar conceptual vocabulary. (i.e. semantically substantive words, see a more detailed description on the generated wordlist below). Our analysis first lists the occurrence of all words in a publication (1). It then groups these publications into clusters based on the co-abundance of conceptual vocabulary (2, 3). In a final step, it identifies (4) representative vocabulary (indicator words) for each cluster (4) and locates these words in two-dimensional space (5). This yields our final word cloud (figure 4). All statistical analyses were carried out using R 3.5.2.

1. ***Digitizing PDFs and metadata:*** To digitize the publications, R imports the 209 PDF files to the working directory and creates a matrix (packages: “snowballC”, “tm”, function: “readPDF”) for further processing. The matrix consists of 209 rows that correspond to number of papers and 20 columns. One column corresponds to the full text of the publication, the others are filled in a next step with general and bibliometric metadata of each publication (e.g., Title, Year, Journal, Citation per Year, DOI, etc.) obtained from the SCOPUS database (code: “scopus.R”, output file: “metaMatrix\_with\_DOI”, both available in SI).
2. ***Wordlist generation:*** To identify the list of conceptual vocabulary, we first generated a complete list of abundant words within the 209 analyzed publications (4720547,205 words), of which 80828,082 words appeared in more than 5% of the publications. Of these, we manually removed all abstract nouns, e.g. pronouns, articles, numbers, authors’ and geographical names, compass directions, units for time, lengths, and mass, as well as individual words with no association to food systems or change processes, or words from which no clear meaning could be inferred. For example, “collect” was retained for its description of a harvesting technique. In this way we retained a list of “conceptual vocabulary” of 2588 words (see Appendix A3b).
3. ***Building publication communities through clustering:*** Based on the co-abundancy of these words, we performed an agglomerative hierarchical cluster analysis using Ward’s method (function: “hclust”, package: “mclust”). This method has been described to cluster “single elements (i.e. publications) into aggregates of two elements based on the minimum variance criterion.” in order to “minimize within-group variance and maximize dissimilarities between groups” (Abson *et al* 2014, p 31). In our case, within-group variance was low if a similar set of words was used in the publications. Similarly, the dissimilarities between groups were high when each community had a distinct set of vocabulary. Our analysis identified five distinct clusters, with an agglomerative coefficient of 0.83.
4. ***Finding representative vocabulary for each publication community:*** To identify words that characterize the differences between the clusters, we used a Dufrene Legend Indicator Species Analysis, which is commonly used in biology “to identify and compare habitats through characteristic species”. As it would for representative biological species, the analysis yielded representative words (indicator words), for each potential publication community, i.e. cluster (Abson *et al* 2014). We show the five most significant indicator words per cluster in figure 4 and provide an extended list of 25 indicator words per cluster in the Appendix (Ax). Based on these indicator words, we were able to identify a hierarchy of publications according to their representativeness of the cluster. The most representative papers most frequently include the most significant indicator words for that cluster.
5. ***Identifying the thematic landscape:*** We used a detrended correspondence analysis to locate the indicator words according to their relative distance to each other (figure 4). In a final step, we inductively identified gradients labels in the thematic landscape of publications. They derived from indicator words and are later refined by the subsequent content analysis (table 2).