Appendix 1

Figures 12-14 illustrate the selection process for the final set of records included in the analysis. The three set of records stems from the keyword-search and were used to feed the *stm* algorithm. As shown in the figures for all the corpora we removed possible duplicates filtering by the DOI code of each publication. We also decided to exclude Review Papers to focus specifically on applied research; however, none were found among the selected records. The resulted sample was of 628 records: 462 records for TFP, 147 to Misallocation, and 19 addressing both TFP and Misallocation. We did not specify an initial search period, and the oldest documents dated from different periods up to 2025.

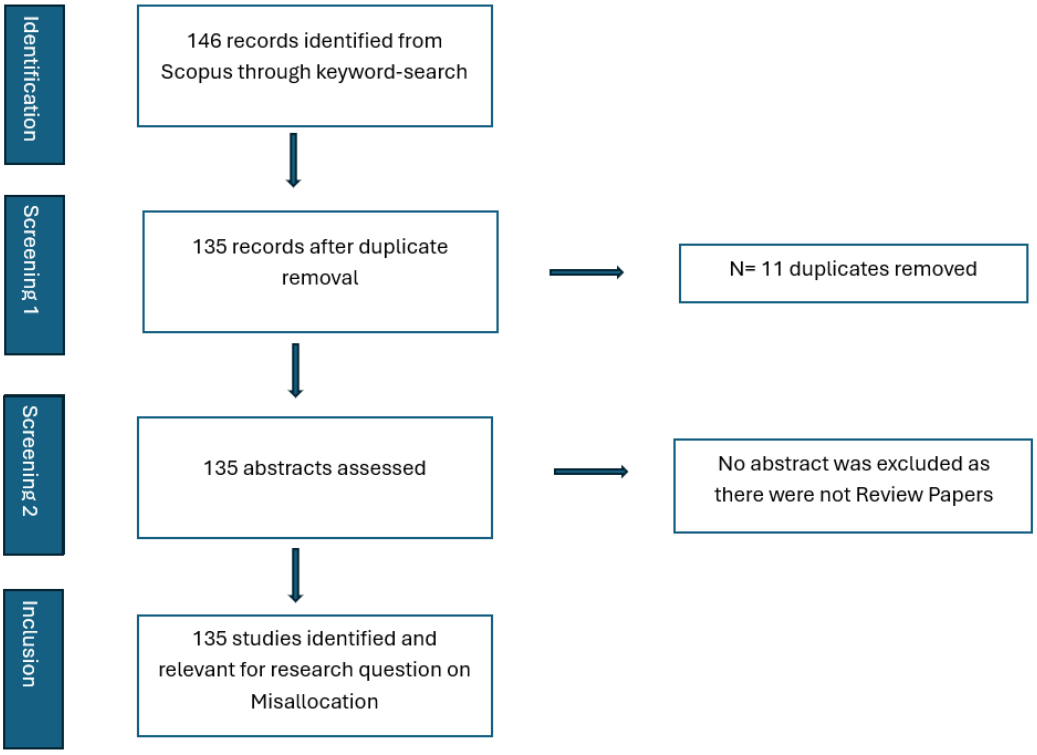

Figure 12. PRISMA flow chart: Misallocation

Source: Authors' elaboration on data from Scopus

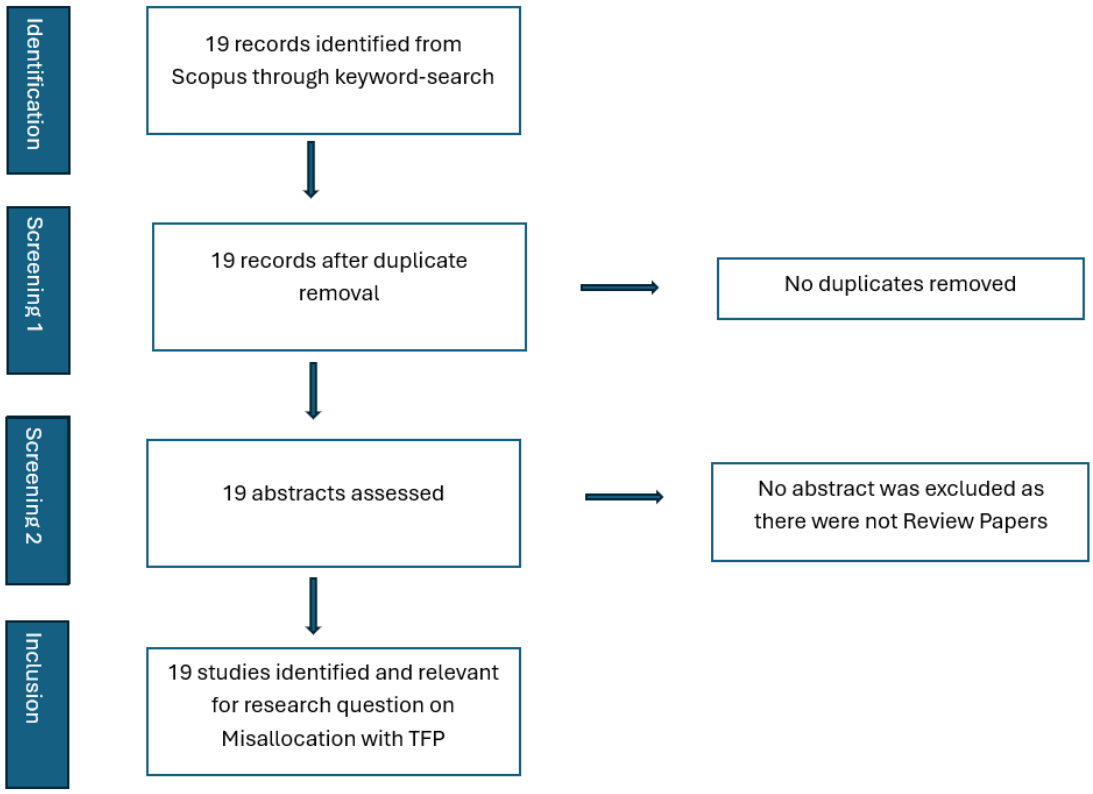

Figure 13. PRISMA flow chart: TFP with Misallocation

Source: Authors' elaboration on data from Scopus

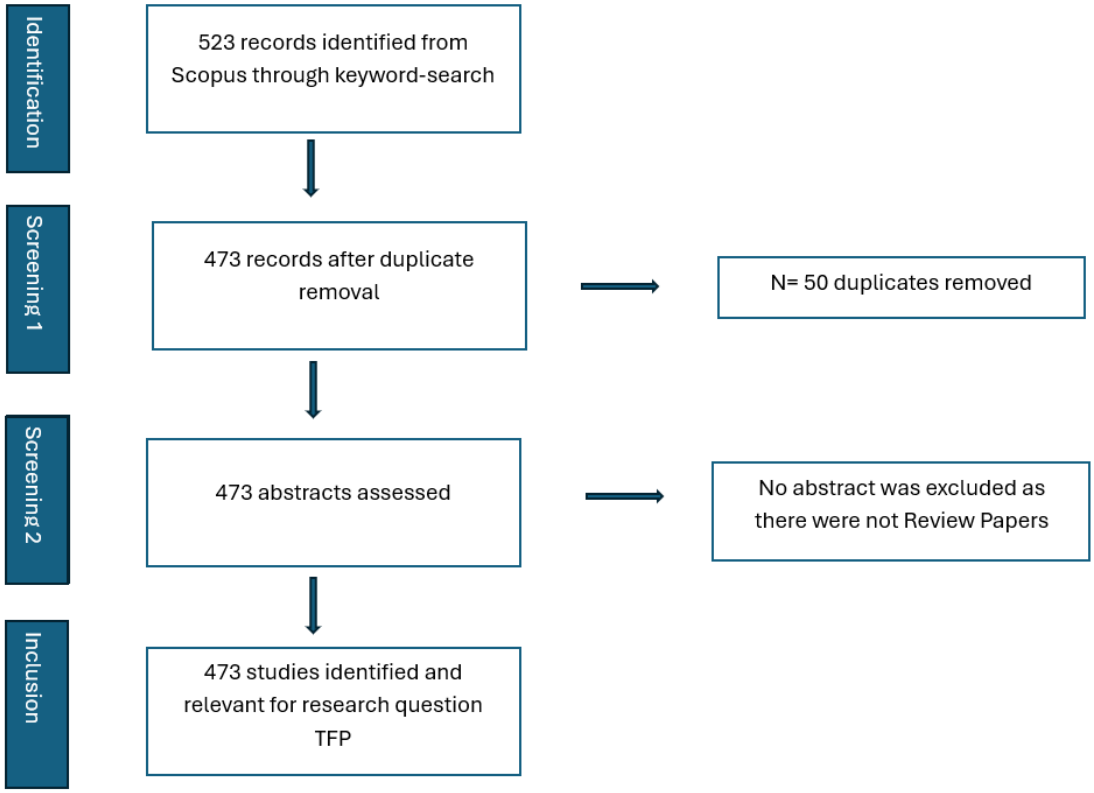

Figure 14. PRISMA flow chart: TFP with Misallocation

Source: Authors' elaboration on data from Scopus

Appendix 2

We fed the *stm* algorithm with the three separate corpora stemming from the keyword-search (see Data Section and Appendix 1 for more insights). To feed the *stm* algorithm we took the *Abstract* of each publication, eliminated all the stop-words, numbers and other recurrent words. Words such as "Productivity" or abbreviations (e.g., TFP, GTPF) were treated as separate terms. However, for the term "Total Factor Productivity" it was not possible to include it in the analysis as the algorithm treats the three words separately. We thus include in the analysis the abbreviation "TFP" and the term "Productivity" which are both good proxies for "Total Factor Productivity". We decided not to perform lemmatization dealing with a token-based representation. This is because the presence of acronyms, domain-specific terms, could not be represented properly in our analysis if lemmatized. We kept those terms with a minimum frequency of 0.5% and a maximum of 99%. Those terms of every *Abstract* of each paper contributed to form a corpus that was used to perform the other steps of the text analysis. The choice of the optimal number of topics coming out of the STM algorithm is based on the qualitative analysis based on how much of the topics could be meaningfully interpreted. Along with this qualitative analysis we also looked at quantitative measures (Weston *et al.,* 2023). *Coherence* provides information on how often features describing a topic co-occur and topics thus appear to be internally coherent. *Exclusivity* is related to how much topics differ from each other and thus appear to describe different things. *Variational Lower Bound* is the metric used to analyse convergence towards a specific solution. *Residual* is the estimation of the dispersion of a given solution. Figures 15-17 shows the different values of those measures according to the different number of topics K for all the three corpora.

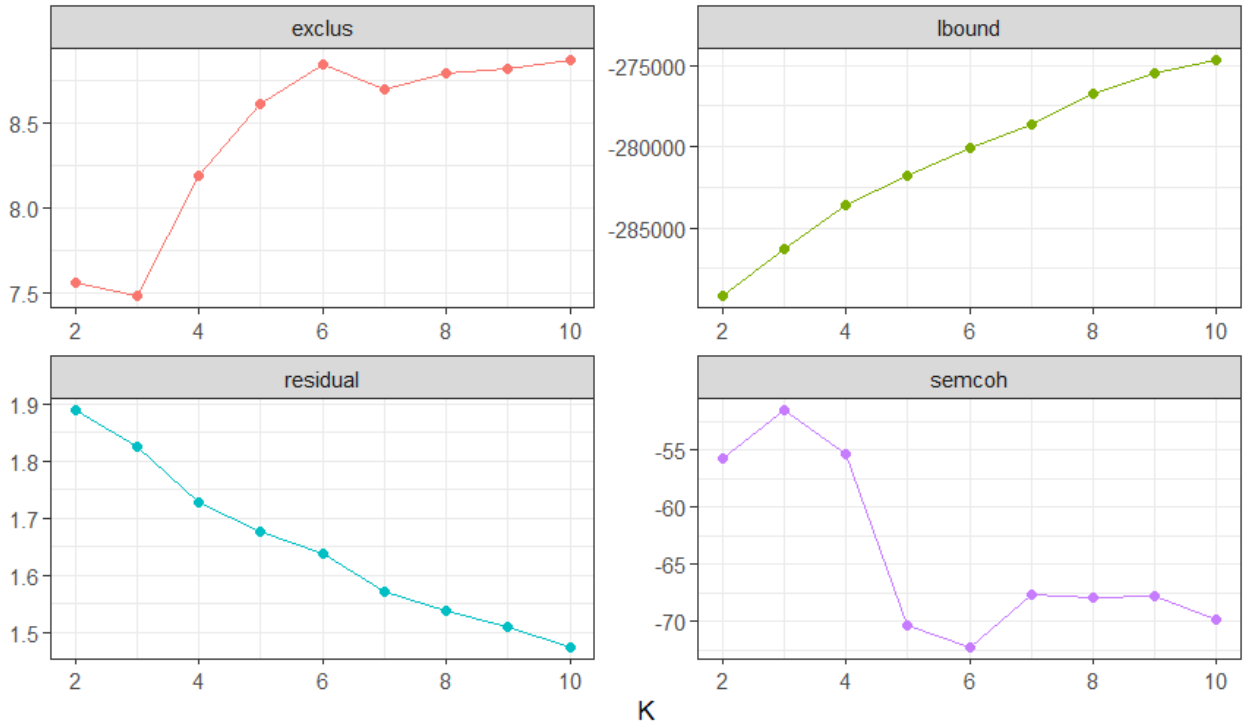

Figure 15. Measures for different values of K: TFP

Source: Authors' elaboration on data from Scopus

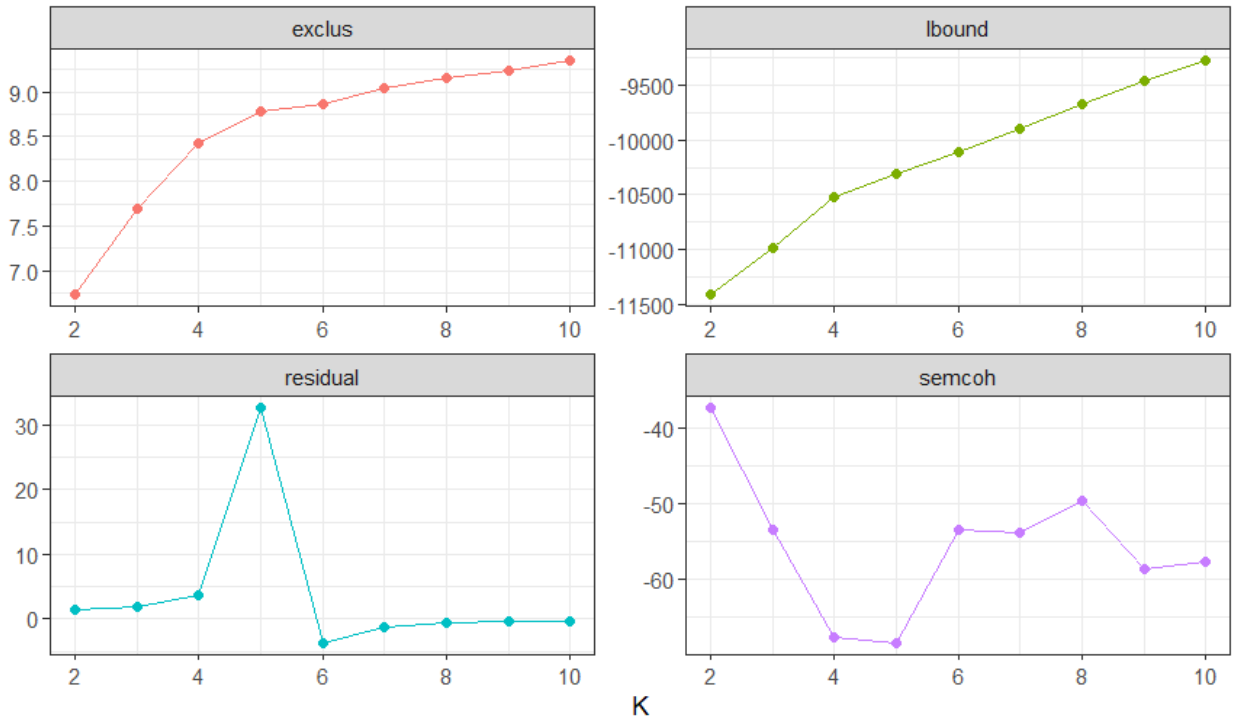

Figure 16. Measures for different values of K: Misallocation

Source: Authors' elaboration on data from Scopus

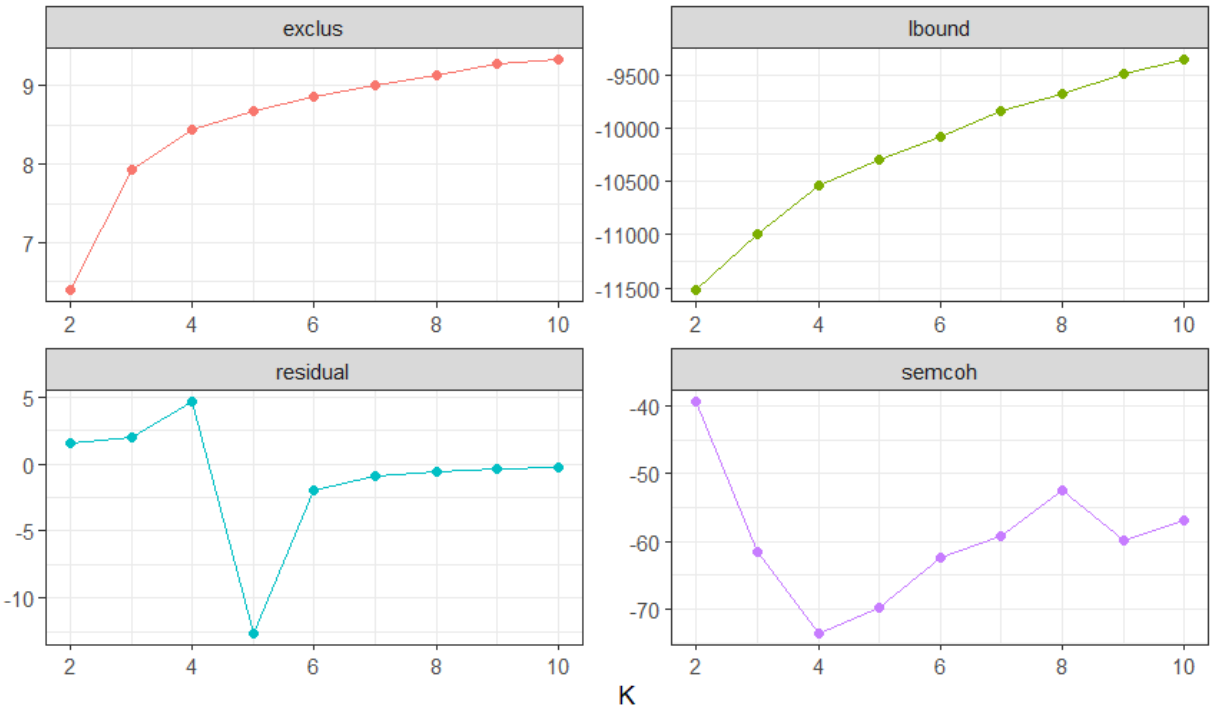

Figure 17. Measures for different values of K: TFP with Misallocation

Source: Authors' elaboration on data from Scopus

Ideally, optimal solutions would be those K for which the model shows higher exclusivity (*exclus*), lower coherence (*semcoh*) along with lower residuals (*residual*) and variational lower bound (*lbound*) (Weston *et al.,* 2023). As the Figures show, considering all four statistics, for the three corpora, optimal solutions lie within the values of K = 4 and K = 6. We finally decided to use K= 5 as the optimal solution for the three corpora because this solution delivered qualitatively more meaningful topics.

Appendix 3

Co-occurrence matrix: Misallocation

Table 7 shows the co-occurrence matrix for the terms "Misallocation", "Agriculture" and "TFP". This matrix is computed from the corpus '*Misallocation*' containing 147 records. This allows us to see if the terms have been used in the same documents. The higher is the number of co-occurrences the higher is the number of papers in which those two topics have been studied together.

Table 7. Co-occurrence matrix for "Productivity", "Misallocation", "TFP": Misallocation

Word	Productivity	Misallocation	Agriculture	TFP
Productivity	57	114	60	3
Misallocation		277	190	57
Agriculture			96	39
TFP				34

Source: Authors' elaboration on data from Scopus

The results show that the term "Productivity" and "Misallocation" appear together 114 times in whereas "Misallocation" and "TFP" 57 times.

Co-occurrence matrix: Misallocation and TFP

Table 8 provides the co-occurrence matrix for the terms "Misallocation", "Agriculture" and "TFP". This matrix is computed from the corpus '*Misallocation and TFP'* containing 19 records. A higher number of co-occurrences indicates that the two terms were studied together in more documents.

Table 8. Co-occurrence matrix for Misallocation, TFP, Productivity, Agriculture: TFP and Misallocation

Words	Productivity	TFP	Agriculture	Misallocation
Productivity	38	39	14	113
TFP		34	3	57
Agriculture			1	24
Misallocation				123

Source: Authors' elaboration on data from Scopus

There are 57 instances where *Misallocation* and *TFP* appear together, and 113 instances where *Productivity* and *Misallocation* co-occur in the same document.