# Meso-level retrieval:
# field delineation and hybrid methods

Michel Zitt
LERECO, INRA SAE2, Nantes, France
michel.zitt@nantes.inra.fr
http://perso.numericable.fr/mzitt/

The meso-level of retrieval (delineation of fields) can be addressed:

By usual IR process and/or bibliometric mapping

Using various bibl. network (citations, words, affiliations, actors/communities) taken alone or combined

# background

# RETRIEVAL AND BIBLIOMETRICS

Garfield’s citation indexing, which connects a new form of indexing and a basis for modern evaluative bibliometrics

Kessler’s discussions of bibliographic coupling as a retrieval method, whilst coupling also appears as a typical tool for bibliometric mapping. Small, Marshakova co-citation and « research fronts » can index citing literature

Shared authors and common tools: Salton, Spärck-Jones, van Rijsbergen, Ingwersen…

# WHAT’S A FIELD?

What is a field (subfield? large research area?)

**epistemology**: knowledge mix: theories, methods, objects

**sociology**: visible or invisible communities of research sharing norms or interests

**politics/ policy**: institutional framework with stakeholders

**info science**: classification categories, retrieval based on key concepts or key authors

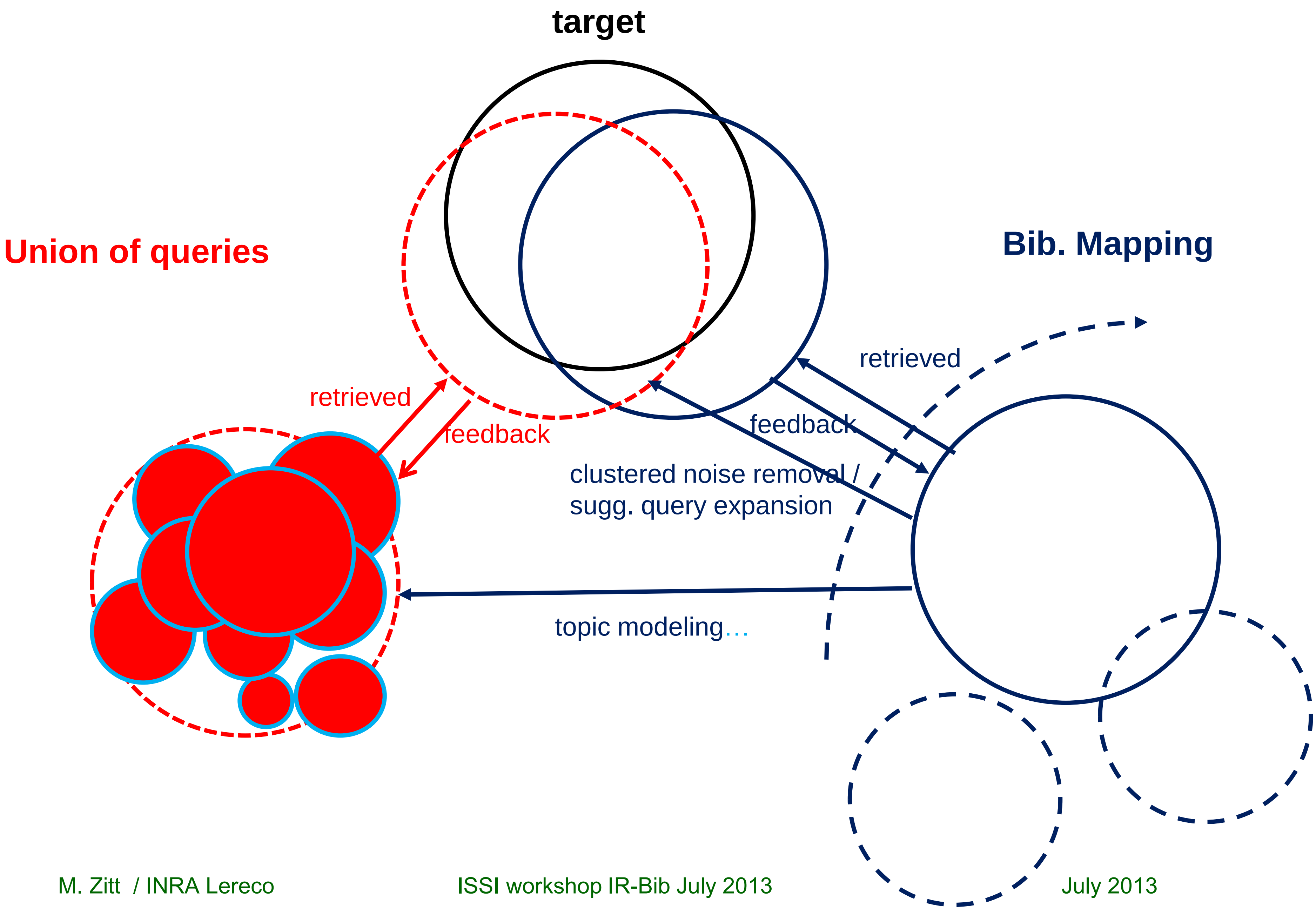
**bibliometrics:** some dense and segregated areas in science network(s), including citation

**information retrieval exercises** have to deal with this variety of vantage points

→ *In scientometrics, the delineation of fields is often used as a baseline (citation normalization, etc.)*

**Query-based, typically bottom-up**

**Map-base, tyically top-down**



target

Union of queries

Bib. Mapping

retrieved

feedback

retrieved

feedback

clustered noise removal /
sugg. query expansion

topic modeling…

# Meso from Micro or - Meso from Macro?

Building up a field through the union of particular queries (terms, lists of journals…)

(+) much better precision and recall potential than macro-queries (titles of fields; list of journals)

(+) internal borders: local robustness of recall due to topics overlaps

*(-) threat on global recall if poor identification of subfields beforehand*

*(-) threat on precision due to polysemic terms, serious problem in a large collection of queries ; hence difficult tuning in order to exploit the recall potential*

*(-) costly supervision*

Building a field by mapping science

(+) visualization of borders and gradients, reduced risk of missing sub-areas

(+) fast methods with little supervision

*(-) threat on precision: most maps are based on holistic metrics (early Martyn's argument against bib.coupling).*

*(-) scale-dependence*

# Gradients and natural borders

# QUERIES-MAPPING MIX

How to find « natural borders » of the field?

Along with various forms of IR adaptive processes - query expansion and feedback loops - a mix of queries and mapping may be helpful.

For complementing IR, additional mapping may help to identify clustered noise (irrelevant subareas) or missing subareas.

Conversely, areas delineated by mapping may be translated through data analysis/ topic modelling or within-field cluster labeling, helping to improve or complete queries.

However, term-based expansion based on mapping continues to face warnings made by Salton in the 60s.

→ ***Multi-network approach*** *may bring effective solutions to field-delineation issues, as well as science mapping.*

# Multi-network approaches

# Hybridization terms--citations

# Words, citations, actors

Early automated IR and word mapping (e.g. Salton in the 60s)

« Citation against words » Garfield's entreprise of citation index, competitor to classical indexes with term-based retrieval. Citation mapping (Kessler, Small, Marshakova, Griffiths, McCain, White)

« Words against citation », the weapon of translation sociology later ANT (UK-France in the 80s), co-word mapping as knowledge representation (Callon, Latour, Courtial, Turner team + Whittaker, Bloor, Law…)

Development of computational linguistics; application of data analysis to texts (CA, Benzecri, 1981). Now LSA, LDA, and other topic modeling methods, most applications on natural language.

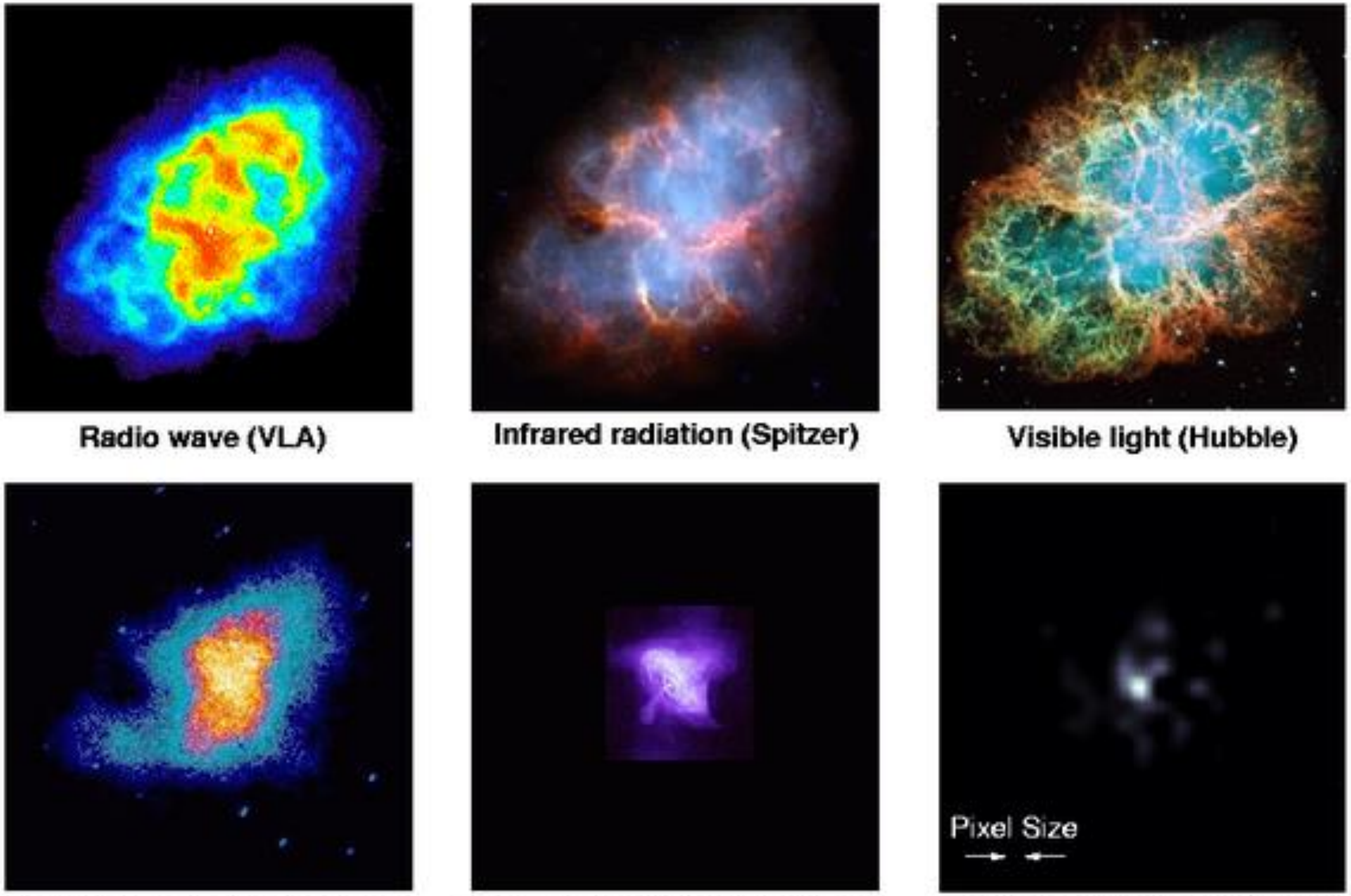
Mapping of actors level (incl. cocitation etc.): perhaps less precise for themes mapping, but powerful in other respects (sociological insights, migration, etc.)

Bibliometrics: pragmatic use of complementarity: adding words to citation indexes (ISI's Keywords-Plus; giving titles to citation clusters)

Citations in context (Small, Attardi, Teufel, Chen…)

# Observing in different wavelengths…

**Crab Nebula: Remnant of an Exploded Star (Supernova)**



Radio wave (VLA)

Infrared radiation (Spitzer)

Visible light (Hubble)

Ultraviolet radiation (Astro-1)

Low-energy X-ray (Chandra)

High-energy X-ray (HEFT)
*** 15 min exposure ***

http://imagine.gsfc.nasa.gov/docs/science/know_l2/multiwavelength.html     Grouping: http://mail.colonial.net/~hkaiter/Telescope_Page.htm

# Word and citation networks

## Bibliometric networks exhibit common formal features…

Originate in the **authors' choice**: composing their text and their references lists, both reflect scientific and social aspects (community markers, enrolment phenomena)

Both are **holistic** from the semantic point of view: regardless to their conceptual status: they mix up theories, methods, applications, etc.

**Formal analogies**: same basic formalism (matrices docs/items and derived, family of skew distributions)

## … but properties are not identical

Citation are **diachronic**, words (primarily) achronic, but bridges can be found

**Statistical properties** are different: word distributions are more concentrated and less « complex », word-relations matrices are less sparse, likely to be more noisy, citations are more prone to clique effects.

**Citation data are less universal**, sometimes difficult to access, but **much easier to handle** and match than texts in natural language (NLT).

**Words directly connect to semantics**(concepts), while citations connect to more complex statements.

**Sociological foundations** are different: communities may be close to each other on words and not on citations, and the reverse may be true  Biases exist for both approaches but quite different in nature: Latourian effects for citation, ideologic marks for words, etc.

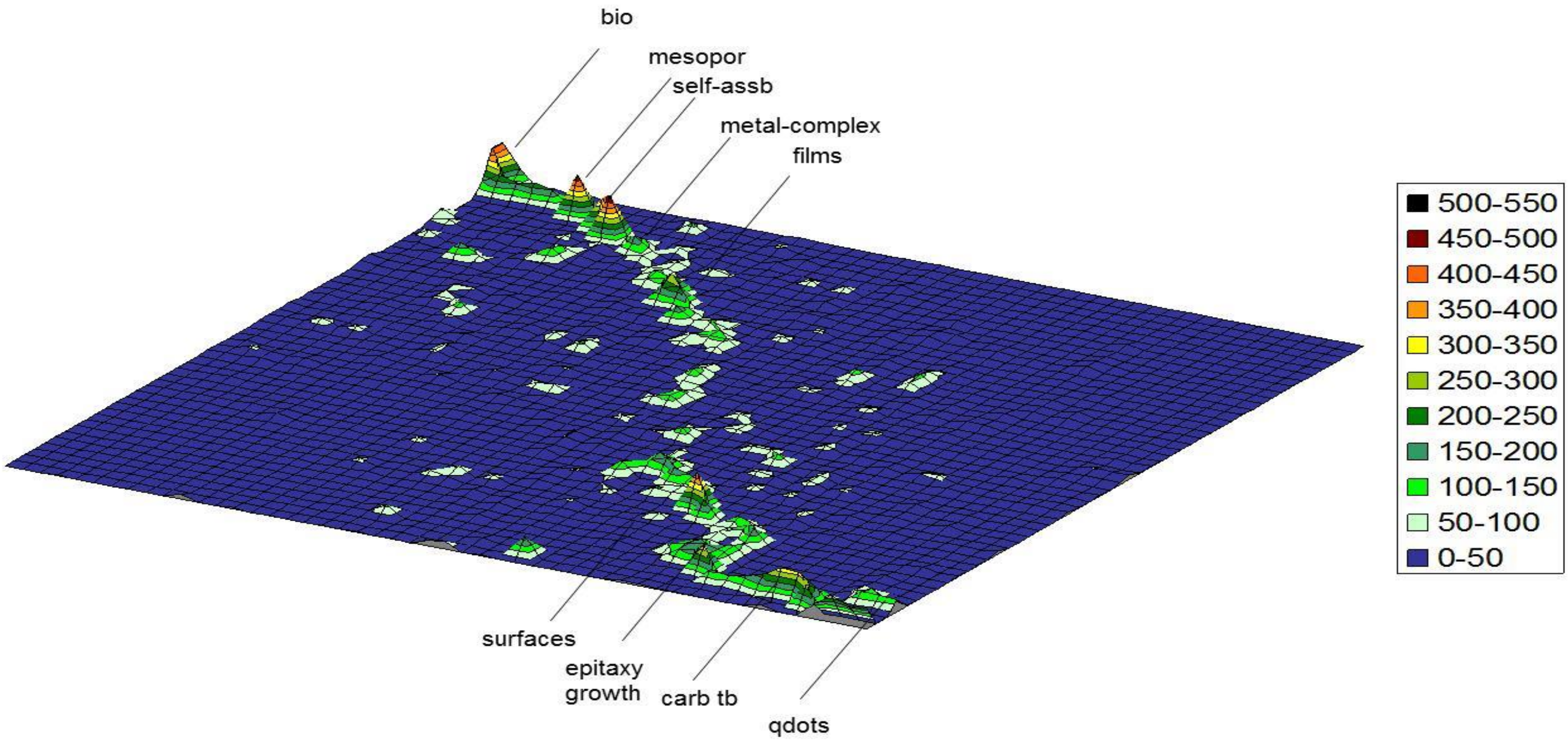
# Citations and words are complementary…

The evidence though mapping: example of cross classification

- same set of articles, two relations (bibliographic coupling, word coupling), same clustering method (variety of k-means, axial k-means), same number of clusters (50)
- process in parallel, the clustering are established separately until the cross-table
- table of intersections and matrix-reordering

Nanoscience archipelago

“where Lilliputians stand”



Method and Material from Zitt, Lelu, Bassecoulard, Jasist 2011

# Citations and words are complementary

The evidence though mapping: Example of cross classification

Same set of articles, two relations (bibliographic coupling, word coupling), same clustering method (variety of k-means, axial k-means), same number of clusters (50)

Process in parallel, the clustering are established separately until the cross-table

Table of intersections and matrix-reordering

Results:

Robustness of bibliometric mapping: the results are by and large convergent. In line with large-scale mapping findings (Boyack, Klavans, Börner; Rafols, Leydesdorff)

However, differences subsist, even overestimated by the cross-mapping.

Results may be used to build strong forms with high precision, retrieved by both methods; or high-recall expansions of word-based clusters or of citation-based clusters; or analyzing cross retrieval indexes.

# One example of protocol: words first

Complex or emerging fields: considering that

- words and citations are complementary.
- starting a multistep process with experts' help is easier with word queries.
- unsupervised procedures are much safer on citations, with proper precautions, than on words.

A simple two-step protocol was set up

- on a **supervised precision-oriented phase:** standard queries: terms, journals → seed
- and a **non-supervised recall-oriented phase:** enhancement by citations → cousin literature sharing the same specific intellectual base

The process is justified by the risks of query expansion on words, already stressed by Salton (1986). The rationale is mapping even if mapping is optional. The citation expansion is relatively safe.

→ *This lex+cite method (Zitt & Bassecoulard, 2008 ) proved efficient for delineation of complex and emerging domains.*

# A variety of protocols

## SERIES HYBRID: citation first

early works on word tagging of cocitation fronts (Small, Griffith…)
extension of co-citation clusters coverage (Braam et al., 1991)
validation of cocitation clusters expansion by textual analysis (Zitt& Bassecoulard, 1996)
textual metrics between cocitation clusters (Boyack & Klavans, forthcoming)

## SERIES HYBRID:  word queries first

Large fields delineation, previous slide

## PARALLEL HYBRID: comparison and final combination

Zitt, Lelu, Bassecoulard, 2008

## FULL HYBRID: combined metrics

structuring/ clustering of fields using a common metrics v d Besselaar& Heymeriks, 2006; Janssens,Glänzel, de Moor, 2008

→ *Not forgetting hybrid processing of hyperlinks by a tiny start-up named Google ☺*

# Conclusion

# Cultural differentiation or pragmatic mix?

The question is asked for the mix of IR and bibliometrics, and perhaps more for the combination of the basic bibliometric networks.

Combined metrics (full hybrid)  implicitly carries a radical informetric or « data-mining » posture assuming that words, citations, authors relations… are information items with reducible statistical differences.

Other hybrid methods can reflect different capabilities of each approach: divergences of citation, textual, author-based methods matter as well as convergence.

Parallel methods keep the principle that clusters based on words and citations are anchored in different sociological foundations. Black-boxes effects may be less severe in these approaches.

## Thanks for your attention