Supplementary Material I: Detailed Delphi method process

Literature Review (Phase 1)

In order to identify indicators relevant to this research, we conducted a literature review in both scientific and grey literature to identify peer-reviewed articles, documents and reports that propose, and review indicators sets for urban sustainability and/or urban food systems sustainability. For peer-reviewed articles, we conducted two Boolean searches in Scopus using the keywords "urban sustainability" AND "indicator*" AND ("system*" OR "framework" OR "set") found in the title or abstract (i) and "sustainability" AND ("indicator*") AND "urban" AND "food"[1] (ii). The search was restricted to peer-reviewed articles published from 2015 onward and limited to English and Spanish (*n* = 702). Spanish was included because it is the authors' native language, as well as given the later applicability test in the selected Spanish cities (see next section), with the aim to broaden the search and explore whether relevant developments existed beyond the more widely circulated English-language literature. The temporal threshold reflects the post-2015 momentum generated by the implementation of the MUFPP. In order to identify grey literature, we followed a similar approach to Merino-Saum et al (2020) to identify relevant documents, utilizing the previous keywords to identify relevant reports from expert groups and institutions including indicator sets through a Google search. This was complemented with backward citation tracking of documents, and through the authors' previous professional experience and networks.

The identified documents were then filtered according to whether they proposed or identified core indicators from previous indicator sets based on the following inclusion criteria:

1. Allow for comparability (not tailored for one specific city/region and drawing on previous international indicator sets).
2. Focus on urban scale.
3. Clear and holistic focus on sustainability (either for cities or for food accordingly), rather than concentrating exclusively on a single aspect, such as economic performance.
4. Comprehensive focus on urban or urban food systems (more than one dimension/process/activity).
5. Applicability to measure sustainability status/baseline.
6. Access to indicators.

Although Spanish was included to broaden the search and capture potentially relevant tools or policy-oriented documents beyond English-language channels, the final selection was determined by the inclusion criteria to ensure consistency across all documents reviewed. The Spanish-language documents identified were thus ultimately excluded during screening, as they were tailored to specific cities or regions and therefore did not meet the comparability criterion for inclusion. The application of the aforementioned inclusion criteria yielded the selection of 38 documents, including those specifically proposing systems of indicators/indicator sets or identifying core indicators by synthesizing indicator sets (see Supplementary Material II). As a result, several of the Indicators identified in these documents were then selected based on their coherence with the SUFOS framework and reference frequency in the identified documents, showcasing broad acceptance of their suitability in the literature, and, if possible, a stronger focus on measuring performance (Béné et al., 2020; González-García, 2019). Indicators were then compiled into the dimensions of the SUFOS framework.

[1] The key words for identifying indicator sets for urban food systems sustainability did not include the work framework given that developing indicator and assessment frameworks for this purpose is still relatively nascent in the literature.

Refinement through city testing (Phase 2)

The identified indicators were tested in five Spanish cities: Madrid, Valencia, Barcelona, Valladolid, and Córdoba, selected following funders' requirement to work in Spain, diversity of locations, and level of development of urban food policies. This first test allowed determining whether issues of data availability would arise in their application, leading to a first reduction of the indicator set. This meant identifying potential data sources for each indicator using local, national and European databases, mainly local statistics offices, the Spanish Statistics National Institute (INE), and Eurostat. As the number of indicators remained high, another round of indicator refinement was conducted. This process was undertaken following a more exhaustive list of inclusion/exclusion criteria identified in previous literature (Béné et al., 2019; de Olde et al., 2017; Halla et al. 2022; Moragues-Faus & Marceau, 2019; Merino-Saum et al., 2020; Papageorgiou et al., 2021; Pülzl et al., 2012; Schang et al., 2021). The criteria considered were:

- Affordability: Indicators that are straightforward and relatively inexpensive to measure were prioritised.
- Clearly defined methodology: Indicators for which data can be collected in standardised methods and thus reliably repeated and compared against each other.
- Comparability and benchmarking: Indicators that were based on absolute numbers were either modified to ensure comparability (e.g., normalised per 100.000 persons) or excluded.
- Conceptual relevance and suitability: Indicators clearly related to the dimensions of the SUFOS framework. Moreover, it was also deemed relevant having at least one indicator for each area (food and urban accordingly) in each dimension.

Delphi survey (Phase 3)

The Delphi technique is characterized as follows: a panel – a group of experts; iterative consultation through two or more rounds usually using a questionnaire; anonymity of participants; and each round being informed by the summary responses of the previous rounds (Jünger et al., 2017). The use of two or more questionnaire rounds with statements on a given issue – either developed by the researcher or by panelists – are devised to form a consensus among selected panelists based on their expertise in the field (Boulkedid et al., 2013). In the case of the modified Delphi technique for the development of an indicator set, statements are replaced by the list of indicators developed by the researchers (Campbell et al., 2002). The panelists rate the preliminary indicators throughout the rounds, with a final face-to-face panel meeting in which the panelists discuss and re-rate each indicator (idem). Introducing a group meeting or discussion has been argued to positively benefit the quality of outcome of Delphi processes (Campbell et al., 2002; Landeta 2006; Venda Nova et al, 2023), allowing the clarification of indicators and addressing the complexity of indicator selection (Khan et al., 2019).

A panel of experts participated in a modified two-round Delphi online survey based on the preliminary indicator set identified through the literature review, with final group meeting to discuss any disagreements on the ratings and validate the final indicator set. 33 European experts in urban food systems assessments, were selected as panelists attending to criteria such as gender balance, diversity of public and civil society actors, and expertise across food system areas The appointment of 20-30 experts is recommended in these processes (Belton et al., 2019). The first questionnaire included the preliminary list of indicators resulting from the literature review organized in the layers of the SUFOS framework (for an overview see the dataset [details withheld for blind review]. Panelists were asked to rate (1 not appropriate - 9 extremely appropriate) each indicator based on two key criteria according to previous studies to ensure the operationalization and adequacy of indicators (de Olde et al., 2017; Janoušková et al., 2018, Lozano-Oloya et al., 2019, Moragues-Faus & Marceau; 2019):

- Actionability: The indicator is straight-forward, relatively inexpensive to measure/obtain, and can be easily adapted to this spatial scale.
- Relevance: The indicator is adequate and appropriate to characterise and measure performance concerning sustainability of urban food systems from a socio-ecological and urban perspective.

Open-ended questions allowed to obtain qualitative feedback on the proposed indicators and elicit suggestions for additional indicators. The questionnaire was piloted beforehand with two experts that met the criteria for panelists, but that were not part of the study, helping enhance the clarity of the survey, as well as its functionality. The questionnaire was developed using the software Qualtrics XM, including a brief introduction explaining the SUFOS framework, and sent via e-mail.

The median was used to determine panelist consensus. Indicators for which the median score was in the top tertile (7-9), and at least 70% of panel ratings were in this range for both criteria, were retained after round 1 (Boulkedid et al.; 2013; Khan et al., 2019). Indicators that reached consensus on one criterion but not the other were included for revision in round 2, as well as those with a median score of 1-3 or 7-9 with agreement of <70% or median score of 4-6 for both relevance and viability (no consensus). Those achieving agreement on their lack of appropriateness (*not* actionable *nor* relevant - Median score 1-3 by at least 70% panelists) were removed. The preliminary indicator set was then updated by adding the additional indicators suggested by panelists and updating/rephrasing previous ones based on feedback.

Panelists that participated in round 1 were sent the second-round questionnaire by email. The questionnaire included the revised indicators where no clear consensus for agreement was obtained along with the new suggested indicators for rating, and indicators for which modifications were suggested. This was accompanied by group results of the first-round ratings of revised indicators, asking the panelists to re-rate each revised indicator based on their individual response as well as the group results. To be included in the final set of indicators, indicators had to have a median actionability and validity rating in the top tertile (7-9) and 75% agreement (Jünger et al., 2017). A group discussion meeting was organized at the end to review the final indicator set. The meeting was recorded and transcribed.

References

Béné, C., Prager, S. D., Achicanoy, H. A. E., Toro, P. A., Lamotte, L., Bonilla, C., & Mapes, B. R. (2019). Global map and indicators of food system sustainability. *Scientific Data*, *6*(1), 279. https://doi.org/10.1038/s41597-019-0301-5

Béné, C., Fanzo, J., Prager, S. D., Achicanoy, H. A., Mapes, B. R., Toro, P. A., & Cedrez, C. B. (2020). Global drivers of food system (un) sustainability: A multi-country correlation analysis. *PLoS ONE*, *15*(4), 1–22. https://doi.org/10.1371/journal.pone.0231071

Boulkedid, R., Sibony, O., Goffinet, F., Fauconnier, A., Branger, B., & Alberti, C. (2013). Quality Indicators for Continuous Monitoring to Improve Maternal and Infant Health in Maternity Departments: A Modified Delphi Survey of an International Multidisciplinary Panel. *PLoS ONE*, *8*(4), e60663. https://doi.org/10.1371/journal.pone.0060663

Campbell, S. M., Braspenning, J., Hutchinson, A., & Marshall, M. (2002). Research methods used in developing and applying quality indicators in primary care. *Quality and Safety in Health Care*, *11*(4), 358–364. https://doi.org/10.1136/qhc.11.4.358

de Olde, E. M., Moller, H., Marchand, F., McDowell, R. W., MacLeod, C. J., Sautier, M., Halloy, S., Barber, A., Benge, J., Bockstaller, C., Bokkers, E. A. M., de Boer, I. J. M., Legun, K. A., le Quellec, I., Merfield, C., Oudshoorn, F. W., Reid, J., Schader, C., Szymanski, E., … Manhire, J.

(2017). When experts disagree: the need to rethink indicator selection for assessing sustainability of agriculture. *Environment, Development and Sustainability*, *19*(4), 1327–1342. https://doi.org/10.1007/s10668-016-9803-x

González-García, S., Rama, M., Cortés, A., García-Guaita, F., Núñez, A., Louro, L. G., Moreira, M. T., & Feijoo, G. (2019). Embedding environmental, economic and social indicators in the evaluation of the sustainability of the municipalities of Galicia (northwest of Spain). *Journal of Cleaner Production*, *234*, 27–42. https://doi.org/10.1016/j.jclepro.2019.06.158

Halla, P., Merino-Saum, A., & Binder, C. R. (2022). How to link sustainability assessments with local governance? – Connecting indicators to institutions and controversies. *Environmental Impact Assessment Review*, *93*, 106741. https://doi.org/https://doi.org/10.1016/j.eiar.2022.106741

Janoušková, S., Hák, T., & Moldan, B. (2018). Global SDGs Assessments: Helping or Confusing Indicators? *Sustainability*, *10*(5), 1540. https://doi.org/10.3390/su10051540

Jünger, S., Payne, S. A., Brine, J., Radbruch, L., & Brearley, S. G. (2017). Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. *Palliative Medicine*, *31*(8), 684–706. https://doi.org/10.1177/0269216317690685

Khan, Y., Brown, A. D., Gagliardi, A. R., O'Sullivan, T., Lacarte, S., Henry, B., & Schwartz, B. (2019). Are we prepared? The development of performance indicators for public health emergency preparedness using a modified Delphi approach. *PLOS ONE*, *14*(12), e0226489. https://doi.org/10.1371/journal.pone.0226489

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, *73*(5), 467–482. https://doi.org/10.1016/j.techfore.2005.09.002

Lozano-Oyola, M., Contreras, I., & Blancas, F. J. (2019). An Operational Non-compensatory Composite Indicator: Measuring Sustainable Tourism in Andalusian Urban Destinations. *Ecological Economics*, *159*, 1–10. https://doi.org/10.1016/j.ecolecon.2019.01.008

Moragues-Faus, A., & Marceau, A. (2019). Measuring progress in sustainable food cities: An indicators toolbox for action. *Sustainability (Switzerland)*, *11*(1). https://doi.org/10.3390/su11010045

Merino-Saum, A., Halla, P., Superti, V., Boesch, A., & Binder, C. R. (2020). Indicators for urban sustainability: Key lessons from a systematic analysis of 67 measurement initiatives. *Ecological Indicators*, *119*, 106879. https://doi.org/10.1016/j.ecolind.2020.106879

Papageorgiou, A., Henrysson, M., Nuur, C., Sinha, R., Sundberg, C., & Vanhuyse, F. (2021). Mapping and assessing indicator-based frameworks for monitoring circular economy development at the city-level. *Sustainable Cities and Society*, *75*, 103378. https://doi.org/10.1016/j.scs.2021.103378

Pülzl, H., Prokofieva, I., Berg, S., Rametsteiner, E., Aggestam, F., & Wolfslehner, B. (2012). Indicator development in sustainability impact assessment: Balancing theory and practice. *European Journal of Forest Research*, *131*(1), 35–46. https://doi.org/10.1007/s10342-011-0547-8

Schang, L., Blotenberg, I., & Boywitt, D. (2021). What makes a good quality indicator set? A systematic review of criteria. *International Journal for Quality in Health Care*, *33*(3). https://doi.org/10.1093/intqhc/mzab107

Venda Nova, C., Ni Riordain, R., Baker, S. R., & Zakrzewska, J. M. (2023). An international Delphi survey and consensus meeting to define the core outcome set for trigeminal neuralgia clinical trials. *European Journal of Pain*, *27*(1), 86–98. https://doi.org/10.1002/ejp.2041