Point-by-point response, Editor Review - Georg Veh

Reviewer Comments
Author response
Text in manuscript

Dear Soraya Kaiser and co-authors,

Thank you for submitting your revised manuscript. We are pleased that you have followed the valuable comments of our reviewers and incorporated their suggestions into your revised manuscript.

However, the editorial review has uncovered additional issues that need to be fully addressed before your paper can be considered for publication in ESSD. Please see the annotated PDF for the editor's comments. We look forward to receiving your revised manuscript in due course.

Best wishes,

Georg Veh

Dear Georg Veh,

We greatly appreciate your suggestions and comments. We have implemented all the proposed grammar, spelling, and synonym changes annotated in your .pdf file. Additionally, we transferred all remaining suggestions and comments to this point-by-point response document and addressed them individually.

Thank you again for your valuable feedback.

Best wishes,

Soraya Kaiser, on behalf of all co-authors

Copied from .pdf:

Line 6: split into two sentences.

Thank you for the suggestion. We split the sentence in two.

––

Line 8: acronym needs to be explained in the text before.

Thank you for pointing this out. We corrected this oversight.

–--

Line 13: Please consider splitting this sentence into two sentences, possibly making the text mining analysis a stand-alone sentence.

Thank you for these suggestions. We split the sentence in two and made the text mining part a standalone sentence.

–--

Line 17: It would be good to add more quantitative content to support the claim of 'high spatial accuracy and completeness'. In essence, 'completeness' is hard to achieve in spatial inventories, as this term hinges on the spatial detail that your inventory can measure. I would avoid this term, if possible.

Thank you for this valuable comment. We quantified the infrastructure mapping accuracy and rephrased the sentence, avoiding the term.

Abstract, line 16: The combination of SACHI and OSM enhanced the detail of the usage type classification for infrastructure from 5 to 13 categories, allowing the identification of elements critical to Arctic communities beyond industrial sites. Further, the new inventory integrates the high spatial detail of OSM with the unbiased infrastructure detection capability of SACHI, accurately representing 94 % of polygonal infrastructure and 78 % of linear infrastructure, respectively.

––

Line 19: ESSD papers should provide the link to the dataset, including citable doi, in the Abstract.

Thanks, we included the DOI in the abstract.

Abstract, line 21: The dataset is available on Zenodo under DOI 10.5281/zenodo.8311242.

––

Line 29: This overview is good, but should also include effects on mountain permafrost.

Thank you for your suggestion. We included a sentence on this.

Page 2, line 31: Permafrost warming trends can also be observed in mountain regions worldwide (Biskaborn et al., 2019), leading to the destabilization of slopes and increased movement of rock glaciers (Haeberli et al., 2013; Haeberli et al., 2024).

--

Line 33: maybe broaden this overview to landslides and rockfalls, as rock avalanches may only refer to the most catastrophic ones.

Thank you. We included rock falls as another type of downward mass movement.

Page 2, line 37: [...] and mass movement processes such as rock avalanches and falls in mountainous regions (e.g. Bessette-Kirton and Coe, 2020; Smith et al., 2023; Stoffel et al., 2024)

--

Figure 01: This is certainly a nice figure, but I wonder whether it makes sense to focus on Alaska explicitly. You mention many locations in the text (Fairbanks, Anchorage, Deadhorse, Homer, Nome, Shishmaref, and so on), and the reader might wonder where these cities are.

I would therefore encourage you to zoom in this map to the state of Alaska and include all locations that you mention in the text. Where appropriate, refer to the figure in the text accordingly.

Thank you. You are right, readers will profit from knowing the location of the settlements. We decided on a pan-Arctic map and an inset map containing the mentioned settlements.

--

Line 79: what 'core categories' are you referring to here?

The core categories, which we are referring to here include food and water supply, banking and finance, government services and institutions, transport and mobility, information and communication, energy production, health and sanitation. We provide a full enumeration in section "Infrastructure and Human-Impacted Areas - Infrastructure Usage Types", page 9, line 209. We included a reference in the introduction:

Page 4, line 81: Following the CIIP manual (Critical Information Infrastructure Protection, CIIP2008) (Brunner and Suter, 2008), we define critical infrastructure as those sectors essential for the reliable functioning of communities. These core categories include among others food and water supply, and health and sanitation. To better align with the modern and traditional ways of life in the Arctic and subarctic region, we have adjusted the internationally recognized core categories and extended them. Please refer to Section 2.2.1 (Infrastructure Usage Types) and Table 1 for a full list of categories.

--

Lin 104-110: I would restructure the first three sentences in this paragraph and start from the most important industries (oil and gas), to fisheries, and then the smaller ones.

Thank you for these valuable suggestions, we restructured the paragraph and added some information on the fishing industry.

Page 5, line 110: The most important contribution to Alaska's economy stems from the mining, quarrying, and oil and gas extraction industry (Bureau of Economic Analysis, 2023a). Notably, the oil exploration units in the North Slope and Cook Inlet play a vital role in Alaska's revenue, having contributed 38 % of the general funds in the 2019 fiscal year (Alaska Oil and Gas Association, 2020, 2021). In addition to the significant impact of oil and gas, Alaska's fishing industry also plays a crucial role in the economy. The Alaska Seafood Marketing Institute (Alaska Seafood Marketing Institute, 2024) reports that, in 2021/22, the fishing industry employed 17,000 Alaskans (from a total of 48,000 workers) from more than 142 communities, making it the top employer in the Alaskan manufacturing sector. Moreover, more than 60 % of the total U.S. seafood harvest comes from Alaska's fisheries (Alaska Seafood Marketing Institute, 2024). Further industries contributing to the economy are: transportation and warehousing (including cargo, passengers but also tourism), finance, insurance, real estate, and government and government enterprises (including community services such as military, postal service, etc.) (Bureau of Economic Analysis, 2023a, b).

However, the economic growth comes with environmental consequence.

--

Line 120: I think you could get rid of Table A1, if you indent i) to v) in the text as separat list elements and include the date of acquisition in brackets.

Thank you for this comment. We agree, the information can be integrated into the text. We followed your suggestion.

Page 5, line 128: The SIRIUS (Synthesized Inventory of CRitical Infrastructure and HUman-Impacted Areas in AlasSka) dataset synthesizes data from five different sources:

(i) Sentinel-1/2 derived Arctic Coastal Human Impact dataset (SACHI) (Bartsch et al., 2021) (acquired June 11, 2021),

(ii) OpenStreetMap dataset for the infrastructure and land use information (OpenStreetMap Contributors and Geofabrik

GmbH, 2018) (acquired January 20, 2023),

(iii) Pan-Arctic Catchments Database (ARCADE) for the watersheds (Speetjens et al., 2022) (acquired January 17, 2023),130

(iv) Modeled Northern Hemisphere permafrost map by Obu et al. (2018) (acquired August 31, 2023), and

(v) Contaminated sites database and reports by the State of Alaska Department of Environmental Conservation (2023a) (DEC) (acquired March 2, 2023).

--

Figure A2 (now Figure 02): This is a really helpful flowchart and should defintively go into the main text.

Thank you for this suggestion. We agree, it is better placed in the main text. We moved it accordingly.

--

Line 133: if these libraries are citable, please add a citation or a link.

Thank you, we added the citations.

Line 133: add citation.

Thank you, we added the citation.

--

Line 135: In this paragraph, I was not sure which parts were already done by Bartsch et al. (2021), and what analysis you added. For example, the 40 km buffer and the label 'Use': was it done by you or by others? Please clarify. One strategy is to use active voice ('We assigned...', 'We added...') whenever you performed a specfic task to highlight your own work.

This paragraph describes the SACHI dataset. The passive voice is used because none of the analyses, such as assigning the attribute "Use," were conducted by us. Instead, these analyses were performed by the authors of Bartsch et al. (2021) following Wang et al. (2021). We switch to the active voice on page 8, line 180, to describe the specific parts of the infrastructure data processing that we carried out.

--

Line 145: which 'both' settlements?

We mean Teller and Nome connected by the Teller-Nome-Highway. We rephased the sentence to clarify this.

Page 6, line 157: For example, the settlement of Nome is assigned the general use category "Mining", with no further distinction, and for the Nome-Teller highway connecting the settlements Nome and Teller, the southern part (Nome) is assigned "Mining", while the northern part counts towards the "Fishing" industry in Teller City.

--

Line 183: you mean here 'unofficial'?

Yes, we meant tags created by OSM users which are not represented in the Wiki. We changed the sentence accordingly.

Page 8, line 197: Some tags were unofficial additions created individually by the OSM mapping community,[...].

--

Line 206-209: I could not follow what you did here. Please clarify. “internally”: what does this mean?

Thank you. “Internally” as in overlaying OSM with OSM. We rewrote this section to clarify:

Page 10, line 221: To address this, we sub-sampled the features with the "fclass" building that hadn't been assigned a usage type yet and "internally" overlaid them with features of any other "fclass" (other than building) that already had a usage type assigned. We then assigned the usage type of the non-building feature to the building feature in the overlapping areas. This analysis revealed that the features with the tag building (e.g. a shopping mall) frequently contain various smaller features and, thus, usage types, such as shops, offices, parking areas, and more. To harmonize this, we aggregated these diverse usage types and assigned the predominant usage type.

--

Line 213: needs to mention that you occassionally compared the two datasets, and that this is a more subjective judgement?

Thank you, we changed it accordingly. We now refer to it as "visually examining" the data/ subsets throughout the text. Examples:

Page 6, line 163: When we visually examined the linear transport infrastructure, we observed some gaps in the data, particularly in settlements:[...].
Page 11, line 231: When visually examining subsets of the SACHI and OSM datasets, we again observed that the OSM data had a higher level of detail.

––

Line 226, "shipping containers": are those assumed to be 'permanent'?

Yes, they are repurposed a lot in Arctic communities: freezer storage, mobile labs, hydroponic, etc. We changed the term accordingly:

Page 11, line 244: ,[...] such as repurposed shipping containers, [...]

––

Line 229: but OSM might have a features that were mapped at a finer resolution?

That is true. We address this matter in the Results and Discussion section. Figure 08b, for instance, shows that the SIRIUS inventory contains narrow residential roads with a width of less than 10m (originating from OSM). These segments weren't mapped in the reference dataset to comply with the SACHI resolution, which led to some false positives. Such discrepancies are inevitable when harmonizing data mapped at different spatial resolutions. We set a threshold of 10 meters as this is the minimum feature size we could reliably identify when merging these heterogeneous datasets. These false positives (fclass = "residential") are addressed in the Discussion:

Page 25, line 503: All the false positive values correspond to narrow residential roads or small paths of the SIRIUS dataset. Although clearly visible in the orthophotos, they were not digitized for the reference dataset because the mapping adhered to the Sentinel spatial resolution of 10m. Including the narrow residential roads and small footways in the reference dataset would have improved the accuracy substantially.

––

Line 250: Switch sentence order: 'We downloaded ... to provide a harmonized ...'

Thanks, we changed it.

––

Line 264: what if several of these keywords were mentioned in this section?

Thank you for pointing this out. We changed “medium” to “media”, as all words mentioned in the section were retrieved as contaminated media.

––

Line 276: the reason for this filtering procedure is not really clear to me.

Setting the cut-off value at a probability of 50% for permafrost occurrence aligns with the definition of the permafrost model domain (Langer et al.; 2023). This threshold provides users with an additional filtering option for relevant permafrost data, allowing for integration with mean annual ground temperature and encompasses the continuous and discontinuous permafrost zone (Obu et al., 2019). We changed the phrase accordingly:

Page 12, line 292: Within our study, we integrated the data on permafrost probability fraction and filtered for raster values where the probability of permafrost occurrence was greater than 50%, complying with the definition of the permafrost model domain (Langer et al., 2023). The filtering step provides users with an additional filtering option for relevant permafrost information, as it allows the integration of mean annual ground temperatures.

––

Line 281: Why didn't you choose a watershed dataset that covers entire Alaska? It is a bit unfortunate that the South-East of Alaska is entirely missing in this database.

We chose this watershed dataset as the SIRIUS database serves as a blueprint for a pan-Arctic database. You mentioned that the outlook was vague about future improvements; this was due to questions of feasibility. We now provide an outlook on upscaling the database to cover almost the entire pan-Arctic region. Additionally, the modeled permafrost occurrence probability (Obu et al., 2019) for the South-East of Alaska shows mostly either a value of 0 (e.g. Alexander Archipelago) or no data (glaciers), which made it less critical for the initial scope of our project.

––

Figure 03: What do IS & HI mean? Please avoid abbrevations.

Thank you for pointing this out. We replaced any mention of “IS & HI” with either “infrastructure and human-impacted areas” or “SIRIUS infrastructure” and modified all figures accordingly.

––

Line 314: I don't understand this statement. Please clarify

Thank you, we rephrased this sentence to emphasize the initial limitation of SACHI and the significant expansion achieved by incorporating OSM data.

Page 14, line 329: . While SACHI only covered the coastal region with an area of 62 km², the incorporation of OSM data has extended the infrastructure map to encompass the entire state, now covering an expansive 640, 593 km².

--

Line 322: rather 'likely' than 'in reality'?

No, sorry, we meant “in fact”. Thank you for pointing out the inadequate wording. We changed it.

--

Line 326: how much? could you quantify how much of more area of the usage type is added by incorporating either dataset into SIRIUS?

Thank you for this question. We included a reference to Table A3 and Figure 6, which gives detailed information on the area of each usage type category under i) SACHI, ii) before the internal overlay , iii) OSM after the internal overlay, and iv) our SIRIUS inventory.

--

Line 337: add the name and provide source for this plugin.

Thanks, we added the source.

--

Figure 04 (now Figure 05): please add a grid with coordinates for orientation in each panel. just label panel c) with 'SIRIUS'? I would prefer to see the captions on top of (or integrated in) each panel, not below.

Thanks for these suggestions. We incorporated them.

--

Figure 05 (now Figure 06): would it be possible to replace the numbers here with the actual name for better readibility?

please reverse the order in the legend to be consistent with the order of the bars.

maybe I missed it, but I still can't find what IS & HI mean.

edit: from the caption, it seems to stand for 'infrastructure and human-impact'. Please make this clear.

Thanks, done!

--

Line 347: please be specific.

Thank you for pointing this out. We provided a more specific explanation.

Page 17, line 363: The resulting SIRIUS infrastructure and human-impacted areas inventory not only represents economic activities, but also incorporates fundamental functions for living, including agricultural areas, commercial and residential zones, recreational spaces, waste and water treatment, and community services.

––

Line 354: I am not sure on how you derived the OA. The overall accuracy is calculated by summing the number of correctly classified values (not rates) and dividing by the total number of values.

In your case, you would need to sum all grid cells where reference and SIRIUS match, and divide by the total amount of cells. Did you do this?

Thank you for your comment. Yes, we calculated the overall accuracy as you described. We summed all the grid cells where the reference and SIRIUS matched (true positives) and divided this by the total number of cells. In the text, it states, "the ratio of correctly classified pixels to the total number of positive and negative pixel values, true and false" which reflects the approach you described. Probably, the word "values" was misleading. We changed it accordingly. The approach is also described in the Method Section on page 11, line 257:

A common metric derived from a confusion matrix is the overall accuracy (OA), the ratio of correctly classified pixels (true positive and true negative) to the total number of pixels (true or false) (Albertini et al., 2022).

Page 18, line 372: The overall accuracy (OA) of the confusion matrix represents the ratio of correctly classified pixels to the total number of all pixels (positive and negative, true and false). The OA of the linear infrastructure data of SIRIUS is 0.5.

–––

Line 355: structure: The OA of the linear … is 0.5

Thank you. We restructured the sentence.

––

Line 374: yes, because the structure is still visible in (maybe older) satellite images, which form the basis of OSM? How many years of decommissioning are required for the infrastructure to be deleted from OSM?

There is no rule on how much time of decommissioning is required for the infrastructure to be deleted. OSM mappers are encouraged to delete objects, when they no longer exist (https://wiki.openstreetmap.org/wiki/Nonexistent_features). However, confirming or refuting the usage of restricted areas, such as a runway within an airport, can be particularly challenging. It's important to note that even though features might be removed from OSM, they remain accessible under "History": www.openstreetmap.org

We also included a section on this in Discussion, see your comment further down.

--

Figure 06 (now Figure 07): maybe add the cell count in brackets? Put the caption for (a) and (b) on top.

Thank you for the valuable comment. We agree that putting the cell count in brackets improves the information. We modified the figure accordingly.

Caption, [...] "The values of the matrices were normalized to the 'true' value, representing the ratio of SIRIUS-mapped features to true features (values [0-1]) " : I can't follow, please explain.

Matrices can be normalized to all, predicted or true labels. We chose the normalization by true labels, as it gives us the ratio of correctly and incorrectly classified grid cells for each true class label - the performance of the SIRIUS map for each true class. We rephrased the sentence to be more precise:

Figure 06, caption: The matrices were normalized to the 'true' value, representing the ratio of correctly and incorrectly (SIRIUS-mapped) features for each true class label (values [0-1]).

--

Line 385: This paragraph needs more quantitative information on how many contaminated sites you detected, in how many (or percentages) of all cases the extraction of the cleanup duration or contaminant worked, and so on.

Line 396: how successfully exactly, and how do you measure this success?

Thank you for these valuable comments. We agree this Section lacks quantitative information. We added the relative and absolute number of entries with zero or negative cleanup days, In addition we emphasized that the performance assessment was done by retrieving a sub-sample of the dataset.

Page 19, line 407: To assess the accuracy, we retrieved a sub-sample of 10 data entries from the dataset, see Table A4. We confirm the successful extraction of dates, following the pattern described in Section 2.2.3. The expressions "Sites Added to Database" and "Sites Closure Approved/ Cleanup Complete" were considered as the first and last action dates, respectively. In cases where no specific expressions were present, the first and last mentioned action dates were used instead. However, in 491 (6 %) entries the cleanup duration was recorded as 0 days (for an example see hazard id 361 of the sub-sample in Table A5), and in 214 (2 %) cases, even negative values were reported. This again points to a heterogeneous approach or methodology used by the agency to input data into the database. In these cases, "Site Closure Approved/ Cleanup Complete" was entered on the same date or even before "Sites Added to Database." For the sub-sample, the retrieval of contaminants was highly successful, as all substances and containment structures listed in the DEC glossary (see Table A6) were found. However, any substances not appearing in the glossary won't be retrieved with our approach. Also, the information regarding the contaminated medium was limited as the DEC rarely provides details in the "Contaminant Information" section of the reports. Consequently, we were only able to derive the contaminated medium for 3321 (39 %) out of 8533 sites.

--

Figure 07 (now Figure 08): color in the legend seems to be different than in the map? Why don't you show in panel B False Negatives and True Positives with the same color key as you did in A?

Thank you for your observation. In subfigure A, we display false negatives and true positives, while in subfigure B, we focus on false positives. This is why we chose to use different color schemes for each subfigure. We have double-checked the colors in the legend and the map, and they do match. We replaced A and B with a) and b) to stay consistent with the other figures in the manuscript. We appreciate your attention to detail and hope this clarifies the presentation choices.

--

Figure 08 (now Figure 09): isn't there a lot of infrastructure present, which is not mapped in the reference dataset?

Yes, you are right. We present this in the Results section of the "Accuracy Assessment":

Page 18, line 389: However, substantially distorting the overall accuracy is the high number of false positives: 568 grid cells showed an intersection with polygonal infrastructure in the SIRIUS dataset (Figure 7b), which was not captured in the reference dataset. 23% of these false positives stem from an overestimation of the airport area in the SACHI and an altogether more generous mapping of the area in the OSM data. The Eastern part of the runway, for instance, appears re-vegetated and allows the conclusion that it is no longer in use, despite being still represented in the OSM data (refer to Figure 9b).

Yet, the highest number of false positives originates from areas affected by human activities represented in the SIRIUS dataset. These human-impacted areas posed a challenge in accurately mapping them for the reference dataset on the basis of the orthophotos alone. Some features, for example a playground, were either not visible or difficult to delineate accurately.

Figure 9c shows an example of a human-impacted area mapped as industrial landuse by the OSM community. While the single storage structures are represented in the reference dataset, there was no indication of an enclosed area visible.

––

Figure 09 (now Figure 10): labels of the grid are missing.

Thank you for pointing this out. We changed the figure accordingly.

––

Figure 08 (now figure 09), caption: what is the threshold to decide, which structures are actively in use and thus enter SIRIUS? In the Arctic, infrastructure might persist for many years to decades after being abandoned, and might therefore qualify as 'contaminated sites'? Please comment on this issue in the text.

Thank you for your comment. Currently, there is no specific threshold to determine which infrastructure features are active and included in SIRIUS.

The SACHI infrastructure is derived through unbiased image classification, with the only constraint being that features must be at least 10 meters in size to be identifiable on Sentinel imagery. In OSM, features are generally included if they exist. If a feature no longer physically exists, users are encouraged to delete it or tag it as "nonexistent" (see https://wiki.openstreetmap.org/wiki/Nonexistent_features). If a feature still physically exists but is no longer in use, users are encouraged to tag it as "disused" (see https://wiki.openstreetmap.org/wiki/Key:disused:%2A).

One option would be to specifically filter for OSM tags "nonexistent" and "disused" to highlight potential contamination sites. This is a valuable idea that we will consider for the next update.

Thank you for your suggestion.

We added a section in the manuscript:
Page 25, line 511: In the case of the polygonal infrastructure for the Shishmaref test area, the SIRIUS dataset achieves a representation of 94 % of all true values. Distorting the overall accuracy are the false positives, approximately a quarter of which belong to the section of Shishmaref's airport runway no longer in use. It is important to note, that OSM encourages users to regularly update features. If a user finds a feature no longer physically exists, they should delete it or tag it as "nonexistent" (OpenStreetMap Wiki, 2024b). If a feature still physically exists but is no longer in use, users are encouraged to tag it as "disused" (OpenStreetMap Wiki, 2024a). In this specific context and considering the potentials of contamination, it could be seen as an asset to have former land usage and industrial legacies represented in the SIRIUS dataset. An interesting approach might thus be to specifically filter for the OSM tags "nonexistent" and "disused" – in the regularly updated and historical OSM database – to highlight potential contamination sites.

--

Line 463: What about using Planet images as a compromise in terms of costs and benefits?

Thank you for this suggestions. We added two possibilities for a compromise.

Page 24, line 484: A compromise could involve using satellite imagery from providers that offer educational programs or discounted rates for researchers, such as Planet's Planetscope with a spatial resolution of 4 m (SentinelHub, 2024). Alternatively, deep learning models could be leveraged to generate high-resolution images from lower-resolution sources like Sentinel-2 (Wang et al., 2018).

––

Line 484: a lot of repetition in this paragraph. consider shortening, and focus instead on how to make the dataset more accurate?

Thank you. We followed your suggestion.

Page 25, line 500: While the linear infrastructure data exhibited low overall accuracy in our Shishmaref test area, about two-thirds of the false negatives resulted from a spatial offset (see Figure 8). Thus, the information of a road's presence is indeed given, but with reduced positional accuracy. This is likely the result of an image offset between the MACS data and the imagery used for mapping the road network in OSM. All the false positive values correspond to narrow residential roads or small paths of the SIRIUS dataset. Although clearly visible in the orthophotos, they were not digitized for the reference dataset because the mapping adhered to the Sentinel spatial resolution of 10 m. Including the narrow residential roads and small footways in the reference dataset would have improved the accuracy substantially. Nonetheless, 78 % of the true road grid cells were accurately represented in the SIRIUS dataset. When accounting for the offset grid cells, this value increases to 92 %. This value underlines the effectiveness of OSM for representing linear infrastructure opposed to SACHI. OSM allows not only a clear distinction between roads and adjacent infrastructure areas but also the inclusion of narrow roads and footways. Looking ahead, it could prove beneficial to integrate official data from local or federal agencies (e.g. Alaska Department of Transportation) to evaluate the comprehensiveness of the OSM linear infrastructure data. Further, incorporating the Trans-Alaskan-Pipeline would provide a spatial context for contamination, oil exploration and transportation data.

––

Conclusions: In most paragraphs, you discuss various ways how the dataset could be improved, but it remains unclear, which of these potential improvements will be actually done, and if so, when. I would welcome a statement on any planned steps in the future, and what you plan to

achieve a version-controlled dataset. Is this just the first release of your dataset and do you plan to continously improve it? How can researchers assist you in making your dataset better?

Line 560: This sentence is mentioned in a similar form at least three times in the manuscript.

Thank you for your valuable comment. We agree that the conclusion was vague about future implementations, which was due to questions of feasibility and funding. We have adapted the conclusion to provide more specific information on our planned steps and timeline.

Page 28, line 581: As we move forward, we have identified several steps to enhance the SIRIUS dataset further. Future updates will incorporate the new version of the SACHI dataset, which was released during the review period of this manuscript. Version 2.0 encompasses i) a refinement of the linear infrastructure features, now distinguishing between asphalt and gravel transport infrastructure, ii) airstrips, iii) humanly influenced waterbodies and reservoirs, and iv) additional regions further inland (Bartsch et al., 2023). The inclusion of water reservoirs affected by human activity is expected to improve the "health and sanitation" category by providing information on water and waste treatment facilities.

Further, an improvement of the text mining approach could be achieved by implementing transformer-based Large Language Models such as GPT or BERT. This could enhance information accuracy, density, and open up new pathways to incorporate contamination-related data from heterogeneous text sources, including online reports, historical documents, and analogue text data.

Researchers and volunteers can contribute to improving the dataset by providing feedback, additional data, or participating in (community) collaborative mapping efforts. The integration of OpenStreetMap into the Land Use / Cover Area frame statistical Survey (LUCAS) framework not only promotes harmonization across international boundaries, but also opens avenues for automated and regularly updated data retrieval through Python libraries like OMSnx (Boeing, 2017). Leveraging crowd-sourced data can encourage future mapping endeavors, including the identification of previously unregistered contamination sources.

We aim to establish the SIRIUS dataset as a foundation for multi-source synthesis and data integration initiatives, consolidating infrastructure, environmental, and health-related information to analyze spatial trends and patterns, with the potential to be upscaled to the pan-Arctic region.

--

Figure A2 (now Figure 02): “MAGT” Spell out fully. limitation means 'clip to state boundary of Alaska'?

Yes, thank you. We changed it accordingly.

--