Clustering solution for flux estimates based on gas bubble stream observations in single beam echo-sounder data using gridded averaging

Knut Ola Dølven

July 12, 2023

Introduction

Accurate and reliable quantification of underwater gas seepage is of great importance for environmental research as well as for monitoring various sub-sea human activities. Our knowledge about the sensitivity and extent of natural seabed seepage is still very limited and even though human underwater oil and gas exploitation shall decrease in the near future, existing infrastructure as well as new industries considering e.g. seabed carbon capture projects need extensive monitoring methods for leak detection. Inversion tools such as the FlareHunter and ESP-3 which makes it possible to extract seabed gas fluxes using the acoustic backscatter data from single and multi beam echo-sounder data are crucial in this perspective since it provides a method which can cover large areas in a relatively short period of time. Single beam echosounder systems are also relavitely cost effective and of limited payload which might make these a a viable option for autonomous vehicles. There are, however, several challenges with the current methodology that can be discussed and potentially improved. In this document, I outline a conundrum and suggest new a solution to the clustering methodology used when extracting flow rates from seep sites using single beam echo-sounder data and the ESP-3/FlareHunter software [Veloso et al., 2015]. Everything outlined here is available as Python (and MatLab soon) code at `https://github.com/KnutOlaD/flare_clustering`, the new clustering method as well.

What is clustering and why do we cluster?

A typical singlebeam echosounder insonofies a cone-shaped volume of water where the horizontal acoustic footprint A_{fp} at a given depth D is given by

$$A_{fp} = \pi \left(D \tan(\frac{\theta}{2}) \right)^2 \tag{1}$$

where θ is the acoustic beam opening angle of the echosounder. For the EK60, which is often used for seep detection and flow rate calculations, the opening angle of the cone is 7^o. The resulting horizontal acoustic footprint of the echosounder can therefore be quite large at typical depths of interest, e.g. at 220 meter the acoustic footprint area is $A_{fp} = \pi \left(220m \tan(\frac{7}{2})\right)^2 = 569m^2$ (see Figure 1a)[1].

[1] The 596 m^2 value does not match with $\sim$ 760 m^2 as calculated in Veloso et al., 2015 using the same input values...

The information used for flow rate estimates in FlareHunter and ESP-3 is the acoustic "target strength" obtained as the logarithm of the summed backscatter cross-section by scatterers in the insonified volume. This data is obtained for each ping and at all depths as given by the vertical resolution Δd of the echosounder. Thus, the insonofied volume here refers to a truncated cone with thickness Δd (see illustration in Veloso et al., 2015). The method assumes that the total backscattering cross-section (TS) is a summation of all the backscattering cross-section of scatterers in the truncated cone -volume, which is, after certain post processing steps, assumed to be gas bubbles. The location of the single scatterers within the acoustic footprint of the echosounder is considered to be arbitrary, thus implying that the TS can be produced by several bubble streams with unknown locations within the footprint. In the case where flare observation samples with sufficiently overlapping footprints are obtained, i.e. as in the top-down view in Figure 1b, clustering is applied with the aim to counteract double counting and maximize probability of a best estimate of the total flow rate. In practice, flares that are clustered are treated and counted as a seep clusters, instead of individual flares and gets assigned a single flow rate value which is calculated from the individual flow rates of the individual flare observations in the cluster.

Vanilla clustering solution

This method is included as the *cluster_flowrate_vanilla* function in the clustering.py script in the github repository (`https://github.com/KnutOlaD/flare_clustering`)

The clustering technique suggested by Veloso et al., (2015) assigns an average flow rate to the total area of flares that are sufficiently near each other to be considered a seep cluster. In Veloso et al., (2015), any flare observation with overlapping acoustic footprint at the seabed is clustered. However, the threshold is often defined such that any flare observations with center distance of 1.8 times the radius (instead of 2 times) of the averaged acoustic footprint (of the two flare observations) are being clustered. Anyway, this way of clustering implies that a flare observation cluster can stretch over large areas and contain hundreds of flare observations (see Figure 13a) in Veloso et al., 2015).

Once clustered, Veloso et al., (2015) obtains the flow rate of the cluster by first calculating the average gas flux per unit area F_{avg} in the cluster as

$$F_{avg} = \frac{1}{K}\sum_{i=1}^{K}\frac{F_i}{A_i}, \tag{2}$$

where F_i and A_i is the flow rate and seabed footprint of flare observation i in the cluster. The total cluster flow rate F_{total} is then calculated by multiplying the average flow rate per unit area by the total area of the cluster estimated using a gridded numerical solution

$$F_{total} = F_{avg}A_{cluster} = F_{avg}N\Delta x\Delta y, \tag{3}$$

where N is the number of grid cells with cell size $\Delta x \cdot \Delta y$ in the cluster area $A_{cluster}$.

Problems with the vanilla solution?

To put forth a small discussion regarding the vanilla solution, let's consider three seep observations with flow rate estimates of $F_1 = F_2 = 2$ and $F_3 = 200$ and the same seabed footprint of $A_1 = A_2 = A_3 = 400$. These seep observations were obtained with 20% shared acoustic footprints between A_1 and A_2 and 20% shared footprints between A_2 and A_3 and were therefore considered a seep cluster instead of individual seeps. This means that their individual flow rates were not considered and a common cluster flow rate was calculated. Following the methodology of Veloso et al., (2015) as iterated above, the total flow rate of the cluster can be calculated to

$$F_{total} = \frac{1}{2}\left(\frac{F_1}{A_1} + \frac{F_2}{A_2} + \frac{F_3}{A_3}\right) \cdot A_{cluster} = \frac{1}{2}\left(\frac{2}{400} + \frac{2}{400} + \frac{200}{400}\right) \cdot 1040 = 265.2, \tag{4}$$

where the cluster area $A_{cluster}$ was calculated as $A_{cluster} = A_1 + A_2 + A_3 - 0.2A_1 \cdot 2 = 1040$ since $A_1 = A_2 = A_3$ and we already knew the amount of overlapping area.

In this case, the vanilla solution obtains a weighted average estimate of the three observations. Since the flow rates are only estimates, this might make sense - however - it is quite obvious that the total flow rate of the cluster is an overestimation. After all, the total amount of gas we have observed in the whole area in individual observations are $F_1 + F_2 + F_3 = 2 + 2 + 200 = 204$, thus we have at least (not counting potential double counted gas) a gas volume of 63.2 that is unaccounted for.... I believe that the issue here is with violated underlying assumptions that are needed to make the vanilla solution valid. I believe a minimum set of these assumptions are

1. All flare observations in the cluster overlap the same amount (otherwise there needs to be some weighing added to account for this in the averaging operation).

2. Deviations from the cluster mean flow rate is due to random processes and/or that the central limit theorem applies.

The first assumption is not violated in our example, but is in practical applications always violated unless there's only two flares in the cluster. The second assumption *is* violated in our case. This assumption is also usually violated in practice. A typical sufficiently large sample size for a random process for the central limit theorem to apply is n∼30. In this perspective, the sample size of a typical cluster is typically too small to assume the central limit theorem.

New clustering solution by gridded averaging

This method is included as the *cluster_flowrate_gridded_averaging* function in the clustering.py script in the github repository (`https://github.com/KnutOlaD/flare_clustering`).

We consider a cluster containing clustered flare observations $k \in [1..K]$. How the flare observations are defined as a cluster, i.e. by what technique, is not important, and can be done by the vanilla solution or by a similar technique (e.g. by a overlapping area threshold).

First, we define a grid covering the total cluster area with grid cells $[i, j] \in [1..N, 1..M]$, resolution $\Delta x = \Delta y = \zeta$ and grid cell center locations (relative to the geographic zero reference) $[i\zeta, j\zeta]$.

Given that each flare observation footprint area has outer bounds defined by center location $[k_x, k_y]$ and radius R_k, we can obtain the vector $\mathbf{I}_k$ containing the $[i, j]$ index pairs of all grid cells within the footprint area by including all $[i, j]$ pairs where $i\zeta < k_x + R_k$ and $i\zeta < k_y + R_k$.

Furthermore, we can approximate the total area A_k of flare observation k by

$$\hat{A_k} = \zeta^2 \sum_{i=1}^{N} \sum_{j=1}^{M} \delta_{i,j}, \quad \text{where} \quad \delta_{i,j} = \begin{cases} 1 & \text{when} \quad [i, j] \in \mathbf{I}_k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and, assuming uniform seepage, we can calculate the flow rate from each grid cell, i.e. the gridded observed gas flux per unit area,

$$\phi_k = \frac{F_k}{\hat{A}_k} \tag{6}$$

where F_k is the estimated flow rate from the flare observation using FlareHunter/ESP-3. The total flux from the flare cluster can then be calculated by

$$\Phi = \sum_{i=1}^{N} \sum_{j=1}^{M} \boldsymbol{\phi} \cdot \frac{\boldsymbol{\delta_{i,j}}}{\sum \boldsymbol{\delta_{i,j}}} \tag{7}$$

where $\boldsymbol{\phi} = [\phi_1, \phi_2, ..., \phi_K]$ is the flux per area of all flare observations in the cluster and $\boldsymbol{\delta_{i,j}} = [\delta_{i,j_1}, \delta_{i,j_2}, ..., \delta_{i,j_K}]$ is the delta function from Eq. 5 for the k^{th} flare observation in the cluster.

Essentially, we average flow rates only where areas of flare observations overlap and do nothing where they don't. This way we obtain the two aims of clustering without letting flowrates observed at one location impact our estimates at completely different locations. In other words: Where we have overlapping samples, i.e. *only where the footprints are overlapping*, we calculate the average flow rate of the overlapping samples. Otherwise we let the individual estimates remain unchanged. Referring to Figure 1b, we average only in the greyed out area, instead of for the combined area of the two footprints as in the vanilla solution. The underlying assumption of this approach (which is also an assumption of the vanilla approach), wich it could possibly be useful to also discuss further, is that seepage is uniform within the area of individual flare observations unless we include information from overlapping observations.

References

[Veloso et al., 2015] Veloso, M., Greinert, J., Mienert, J., and De Batist, M. (2015). A new methodology for quantifying bubble flow rates in deep water using splitbeam echosounders: Examples from the Arctic offshore NW-Svalbard. *Limnology and Oceanography: Methods*, 13(6):267–287.

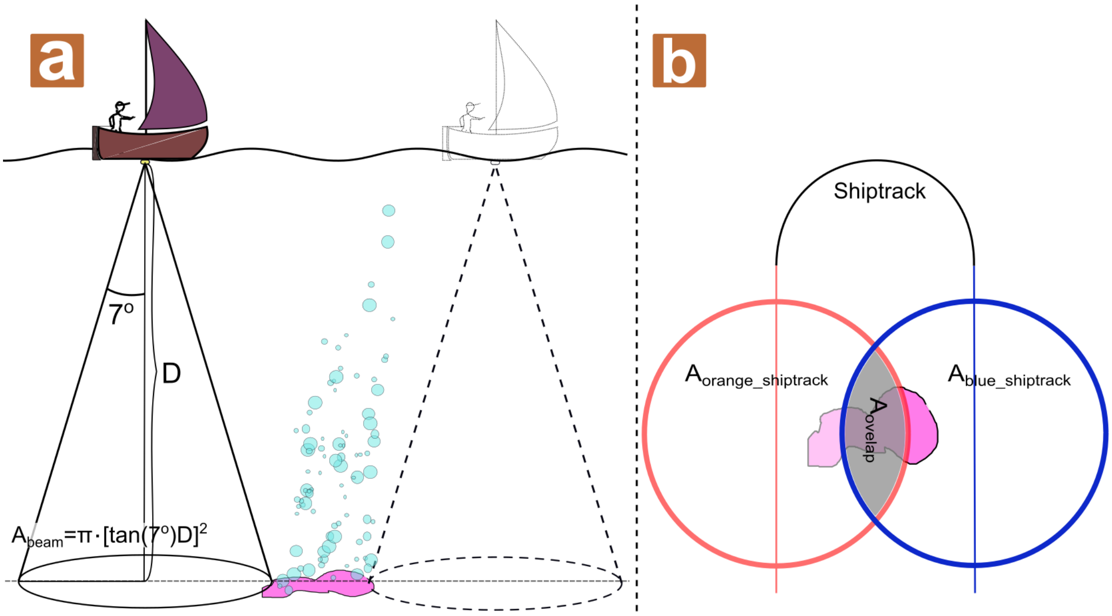

Figure 1: Conceptualized figure showing a) Insonified volume during an echosounder survey and observation of seabed seepage here as a constrained area of seepage and b) top down view of the acoustic fooptint illustrating double counting of the same seepage area and an over estimate of the total flow rate during an echosounder survey.