**How to choose your threshold for neighbour distances**Jessica E. Forsyth

We have developed IVEN as a flexible framework that can be applied to other model systems and datasets. We therefore hope to encourage appropriate manipulation of parameters and methods of analysis.

This tutorial focusses on how to choose what distance threshold should be applied when identifying neighbours of cells, and tips on how to decide what threshold you should choose, what method and whether you need a threshold at all.

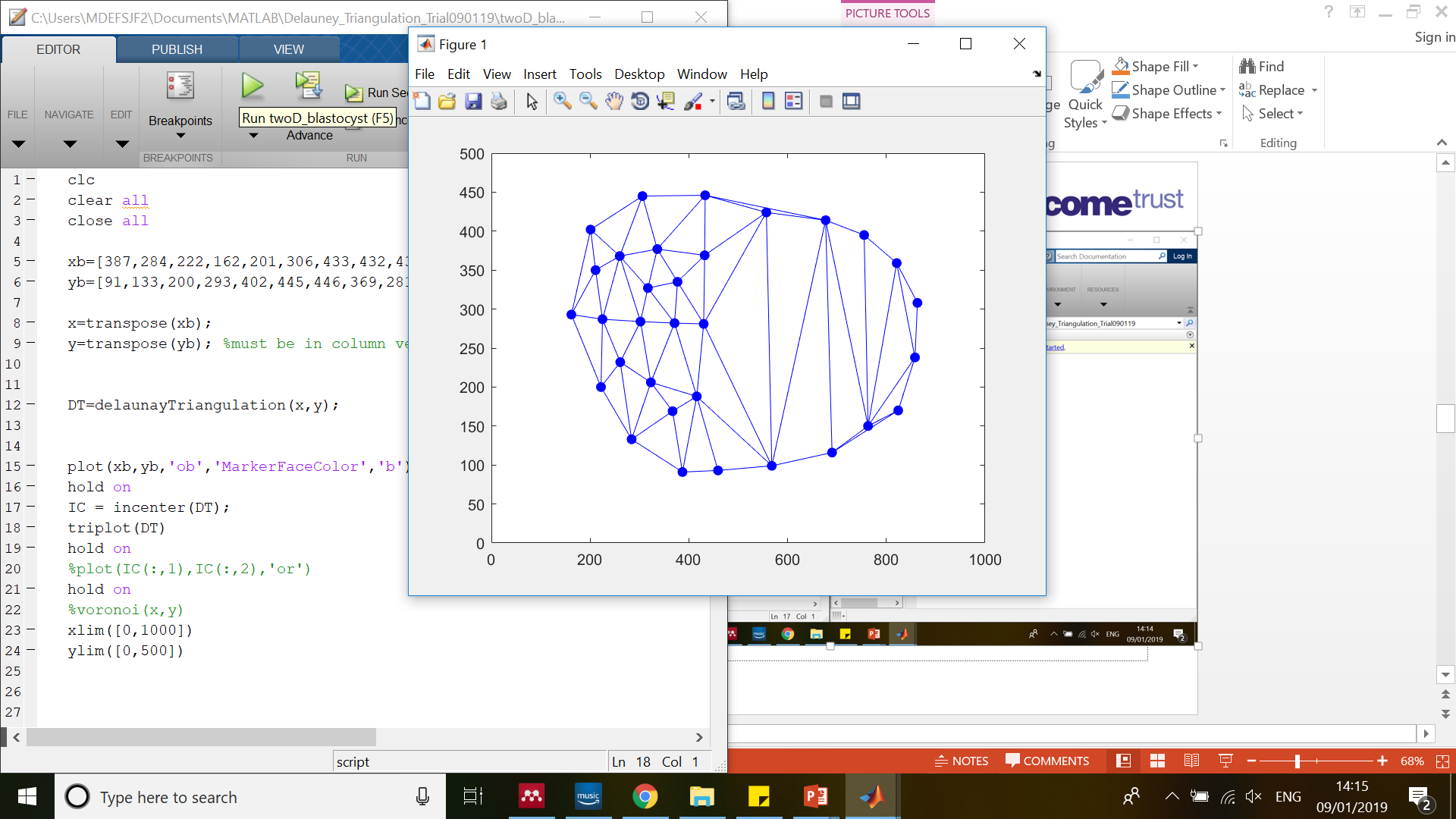
**What do we mean by a ‘neighbour distance’ or ‘distance threshold’?**

Figure Example cross section of a blastocyst with cavity present. Cells on the mural TE are shown to be neighbours with each other (joined by line) depsite being across the cavity from each other.

First when identifying the neighbours of a cell, we generate the Delaunay Triangulation (DT). The triangulation describes the ‘connections’ between neighbouring cells and we use this to infer which cells are neighbours. However in some systems, for example the preimplantation mouse embryo, there are cavities present within the structure. Such cavities are not accounted for during the construction of the DT, meaning some cells are considered neighbours despite being on opposite sides of the cavity (Figure 1).

**Why use a distance threshold between neighbours?**

To avoid this problem, we set a maximum allowed distance between neighbours (‘maximum neighbour distance’ or ‘distance threshold’). We then check the distance between neighbours. If the distance between two cells (originally identified as neighbours) is greater than the threshold distance, we remove these cells as neighbours. By doing this we hope to remove ‘untrue’ neighbours, such as those on opposite sides of the cavity in the mural trophectoderm (TE).

**Why do we need to be careful in how we choose the threshold?**

The threshold distance chosen inherently affects the number of neighbours calculated for each cell within the sample. If the threshold is chosen to be too small, then too many cells will be removed as neighbours (including some true neighbours) and if the threshold is chosen to be too large, then not enough cells will be removed as neighbours (allowing some untrue neighbours to be included).

More specific to this example and application of IVEN- the murine preimplantation embryo, we show that cells undergo cleavages during this period (the cells continue to decrease in size) and so the applied threshold must also adapt to the changing morphology of the embryo itself. Similar changes will be evident within other systems.

So any threshold chosen must be appropriate for not only the model system being used, but also the developmental stage.

**Our (automatic) approach to choose a threshold-**

We have chosen to take a more general approach to threshold our neighbour distances, which is hopefully appropriate to other systems, so long as it is tuned correctly. This approach calculates the threshold from the input data, meaning the threshold is sample specific and therefore aims to remove effects of stage and model system.

1. Calculate the distance between all cells and all of their identified neighbours (using DT).
2. Plot the distances calculated as a distribution.
3. Calculate the 75th percentile of the distribution, and the inter-quartile range.

In most statistical packages, when generating a box plot, outliers are identified as any points above the value , where is the 75th percentile, is the inter-quartile range and is a parameter set by the user.

We also use this approach for a cell type dependent threshold, by calculating the distance distributions between neighbours that are inside cells and those which are outside cells. We then find thresholds for these separately.

We use this same method and choose an appropriate value of for our data at all stages. In order to choose an appropriate value of we performed two analyses;

1. Measure the maximum distance present between two true neighbours (in our case this maximum distance often occurs between mural TE cells as these cells are more stretched over the cavity). We measured this distance for multiple embryos within the 32, 64 and 128-cell stage.   
   We then compared these distances to the neighbour distance distribution plots and identified what value of would permit such neighbours to be retained without allowing untrue neighbours (such as those across the cavity) to be retained.
2. Perform a manual neighbourhood assessment (using nuclear markers and cell membrane markers) for a subset of cells within embryos at different stages. Then calculate the number of neighbours using IVEN for different values of and see which value of gives the best agreement with the manual assessment of neighbourhood.   
   *NOTE: assessing the neighbourhood manually can be more difficult than anticipated so allow for some human error during this process.*

**An alternative approach-**

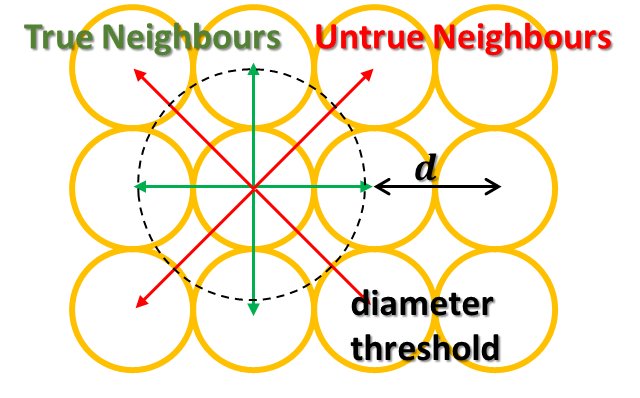
Another method to determine a threshold for your data is to measure key features of your dataset and use this as an input threshold. For instance if your sample consists of highly uniform cells with the same diameter it would be fair to apply a threshold approximately equal to the cell diameter (Figure 2). It would be advisable to impose some kind of threshold to account for subtle differences in packing and/or cell shape.

Figure 2 Two dimensional schematic showing the use of the threshold as the cell diameter in a uniform array of cells. This schematic would also serve in a three dimensional system.

Or if there is a key feature within the sample that should not be included, e.g. a cavity, an approximately of the cavity size could be included in some calculation of the threshold.

This method of choosing the threshold should be informed by the system being analysed and justified through manual checks, however, it is just as valid as the first approach.

**Do you even need a threshold?**

Say you have a highly uniform data set, with no cavities or different features, you may not need a threshold. The threshold is simply employed to avoid artefacts generated from non-regular features of the dataset (i.e. a cavity) or features that disrupt the DT.

Overall the threshold should be evaluated for the specific system being analysed and should be chosen appropriately such that (if possible) all data is processed in a similar way to allow comparison across data sets.