# DC1 all tasks

## Session 2/Part 1

Task 0: Create a new R Project and new R Script using a clean template and initialization of the environment. We recommend you use RMarkdown with HTML as target output.

Task 1: Download and import the road accident data from Open Data Zurich (ODZ) using an appropriate file format. If needed, transform to the Swiss projected CH1903+/LV95 coordinate system (EPSG = 2056).

Task 2: Report the following numbers in table(s):

1. Number of accidents by accident severity category
2. Number of accidents by accident type
3. Number of accidents involving pedestrians, bicycles, and motorcycles, respectively. And combinations thereof (pedestrian AND bicycle, pedestrian AND motorcycle etc.). Are there any accidents involving all three modes (pedestrian, bicycle, motorcycle)?

Task 3: Generate a plot showing the temporal evolution of the number of accidents from 2011 to 2023. Label each year with the corresponding number of accidents. Choose a plot type that is suitable for this type of temporal data. Bonus: Show also the data for the bicycle accidents

(cf. Task 4) in the same plot.

Task 4: Select only those accidents that involved a bicycle. From now on, and for the

remainder of DC1, we will restrict our analysis to the accidents involving bicycles.

With this subset, produce a map showing the bicycle accident data colored by accident severity category. Use a basemap such as OpenStreetMap and/or the boundary data available on OLAT, so the accidents can be visually and spatially referenced.

Task 5: Imagine you are given the task of detecting spatial clusters of elevated bicycle accident occurrence (without considering their severity). How would you characterize such “bicycle acci- dent clusters”? Try to define properties that can be used to describe and identify such clusters, and that can be used to choose and parameterize a clustering method suitable for the task. Try to use natural, but precise and concise language in your answer.

Task 6: From the bicycle accidents, extract the years 2018 to 2021 and compute clusters for each year separately, using a clustering method you deem appropriate for the task, and choose the control parameters appropriately to capture the types of clusters you had in mind in your definition of Task 5. Justify your choice.

Task 7: Discuss your results, including also limitations or problems, and possible other methods that you could have used.

## Session 3/Part 2: Polygon delineation

Note: As in the preceding Part 1 of DC1, we will restrict our analysis to accidents involving bicycles.

Task 8: Given the clusters that you have extracted in Part 1 of the DC1 assignment:

* 1. Define a set of criteria that a method should fulfill that can be used to delineate the given clusters by polygons. Use free text for these definitions, but try to be concise and precise. (Note: These criteria can also be used in the subsequent Discussion to evaluate whether they have been met.)
  2. Choose a polygon delineation method that you deem appropriate in light of the above criteria. Justify your choice.

Task 9: From the years 2018 to 2021 for which you computed clusters in Task 6, choose at least two years and apply your polygon delineation method of choice to each of these two years separately. Compute the Jaccard Index (aka Intersection over Union) for pair(s) of selected years and present and discuss the results.

Task 10: Overall, what did you find with the above steps? What do these steps tell you about the situation of bicycle accidents in Zurich? How useful are the methods used so far in analysing the given data? Any other points of note?

Note:

• As always, visualization helps with both developing your own trains of thought as well as the discussion – even if we don’t specifically mention visualization in the assignments. (Initially, you might even do that in QGIS, if you don’t feel at ease yet programming in R.)

• Overall (i.e. over all assigments), in DC1 20% of points will be awarded for creativity and quality of the approach (structure, depth, presentation). Sometimes, with little additional effort, added value can be generated.

## Session 4/Part 3: Density estimation methods

Note: As in the preceding parts of DC1, we will restrict our analysis to accidents involving bicycles.

Task 11: Similarly to the clustering and polygon delineation tasks carried out in Parts 1 and 2 of DC1, respectively, start off by defining criteria for using KDE to detect areas/hotspots of elevated bicycle accident density, and explain your reasoning.

Task 12: Choose any two years from the years 2018 to 2021 (justify your choice of years) and compute the KDE surfaces for each of these two separately and visualize your results. You are free to choose the KDE implementation (i.e., R package and function(s)) as well as the parameters (bandwidth selection method, etc.), but you should document your choices and discuss, in the subsequent Task 14, your results in light of your choices.

Task 13: Compute the “volume of intersection” (“VI”) between the KDE surfaces (utilization distributions) of the two years. Hint: There are different ways to do that, but the adehabitatHR package has functionality for that. How do the results correspond to those of Task 9 (Jaccard Index or IoU)?

Task 14: Discuss your results for this part of DC1 (density estimation). What did you find? Compare the results of this part with the clusters/polygons of Parts 1 and 2 (see note below): What are the commonalities? What are the differences? Which method(s) perform more adequately than others for the given problem and data? Which method(s) would you recommend, and which ones not? Why? (You are free to add more points to the discussion.)

Note: in order to compare the clusters/polygons of Part 2 with the density hotspots of today’s assign- ment, you may want to overlay these layers in a joint plot. To avoid re-running of large portions of the script of Part 2, you may save the cluster polygons that you delineated in Part 2 using sf::st\_write() and read them from file in Part 3 using sf::st\_read() (if you choose to have separate scripts for Parts 2 and 3). Alternatively, you could use the source() function. See the hints on Slide 2 of Session 3 (last week).

## Session 5/ Part 4: Second order properties

Task 15: Choose one or more distance measure functions (justify your choice) and compute it/them for

1. all bicycle accidents (2011 - 2021)
2. bicycle accidents for one of the years selected in Task 6
3. bicycle accidents for only the cluster points of the year selected in (b)

Restrict your analysis to the “inner city” and use the same window for all point sets. It’s up to you to define the extent of the “inner city” and explain/justify what that means in terms of this data challenge.

Task 16: Now choose the following two pairs of years, 2018 & 2019, as well as 2018 & 2021, and compute the cross-X function, where “X” stands for the function(s) you used in Task 15. Use the AccidentYear as the marks to produce a marked point pattern.

Task 17: Discuss your results. Similarities and differences between the various point sets? Noteworthy spatial and/or temporal patterns? Is any difference observable in the patterns of the cross-X function in the transition to the Covid-19 pandemic? etc. etc. Note: Consider in your interpretation that the distance scale may change between years.

Task 18: From the bicycle accidents data, choose at least two relevant variables that make sense being compared, e.g. two different accident types, severity levels, times of day, years etc. For these selected accidents, compute the counts within the ‘statistical zones’ of Zurich and use these counts to compute the Getis-Ord G\*-statistic for each of your counts layers. Visualize your results appropriately.

Task 19: Discuss your results. What did you find regarding the hot and cold spots in your acci- dent count layers? How do they compare to each other across layers and across (past) methods? etc. Discuss also the influence of the parameter settings (e.g., neighbor search distance, spatial weight formation) on your results.