Algorithmic Data Analysis

Esther Galbrun

Spring 2024

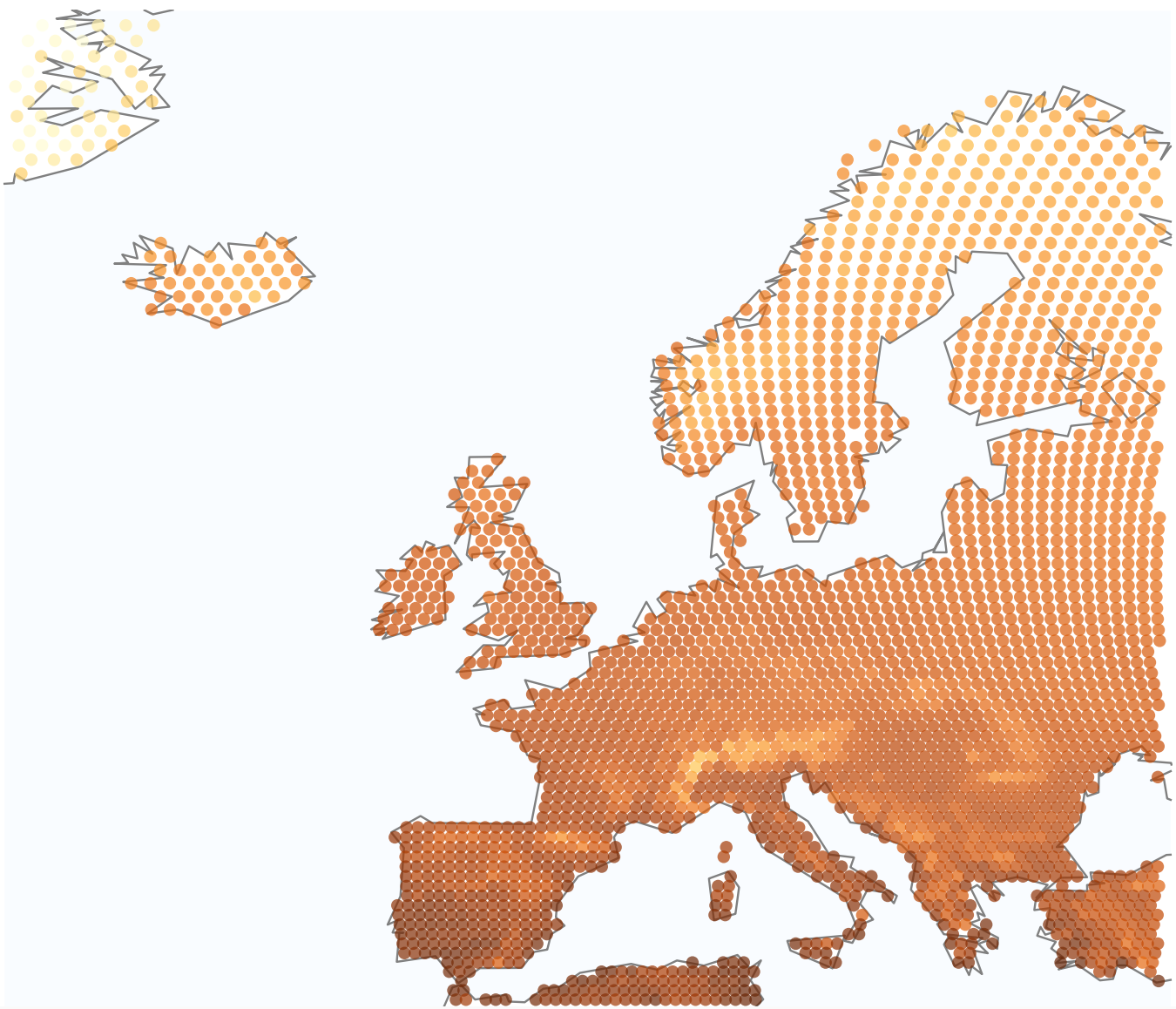
UNIVERSITY OF
EASTERN FINLAND

Last updated on 2023-12-28 11:28

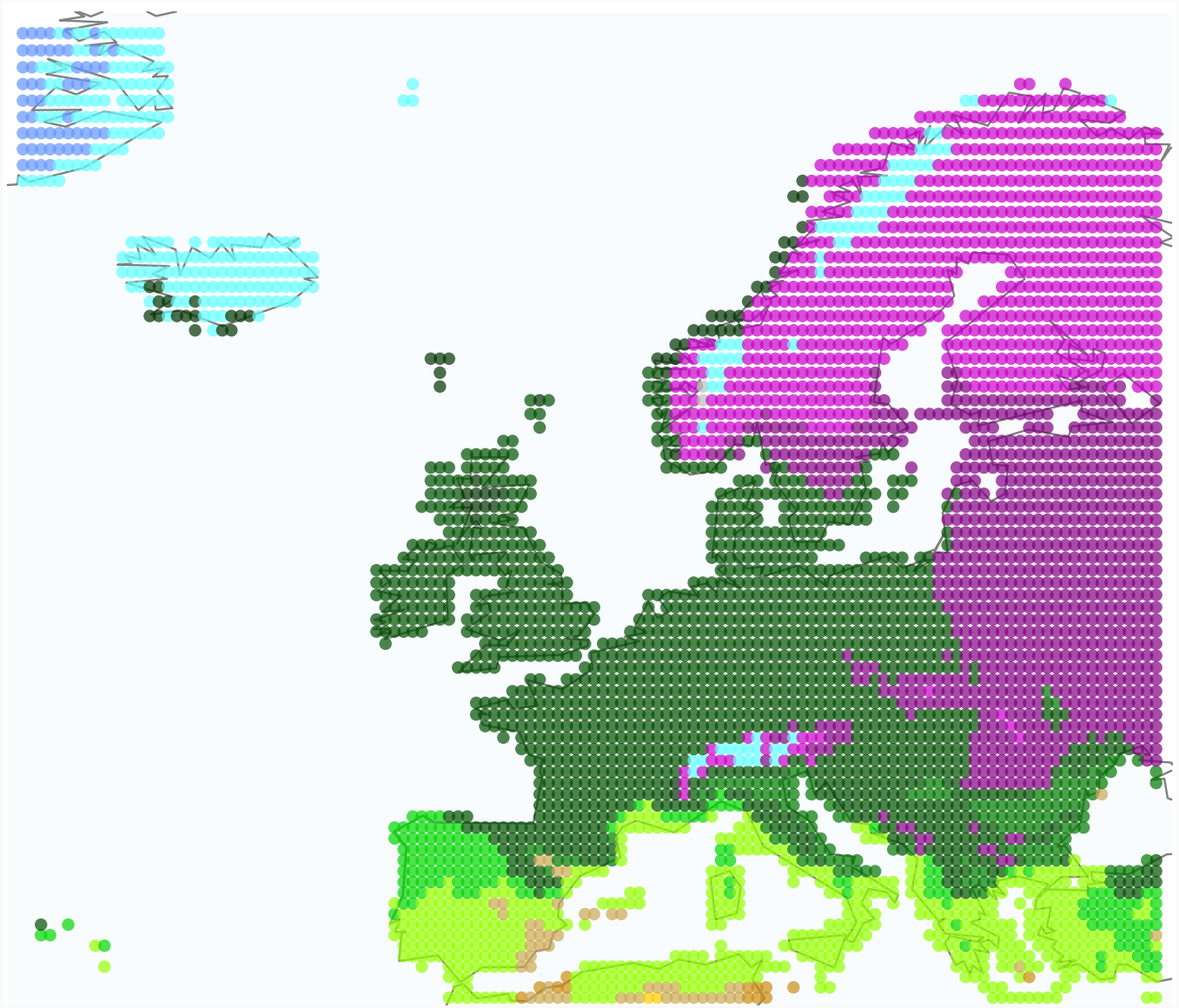
Part VI

Mining spatial data

Yearly mean temperature

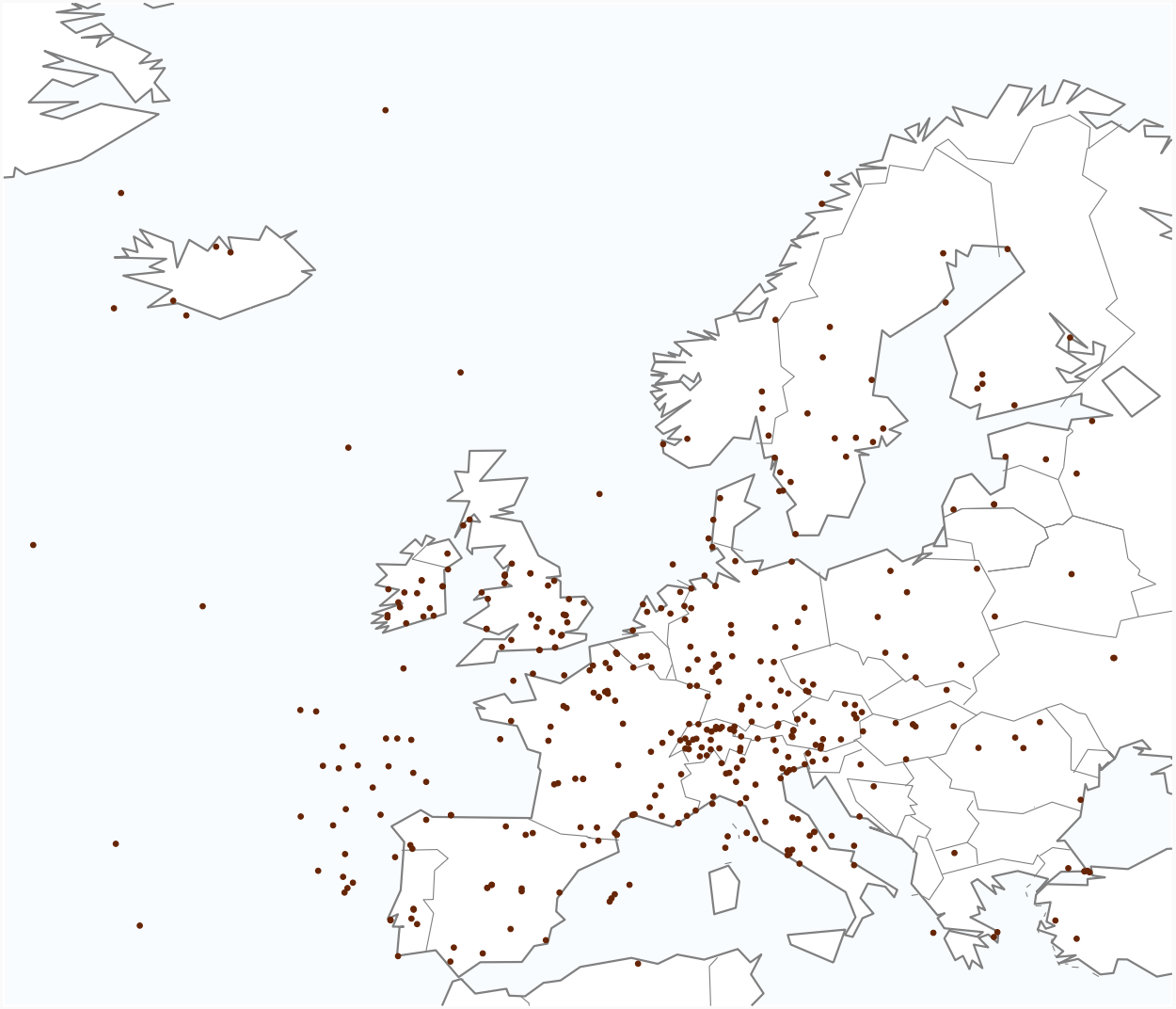


Köppen climate classes



Trees on the streets of Manhattan (NY 10003)

Aviation accidents



Birds migration trajectories

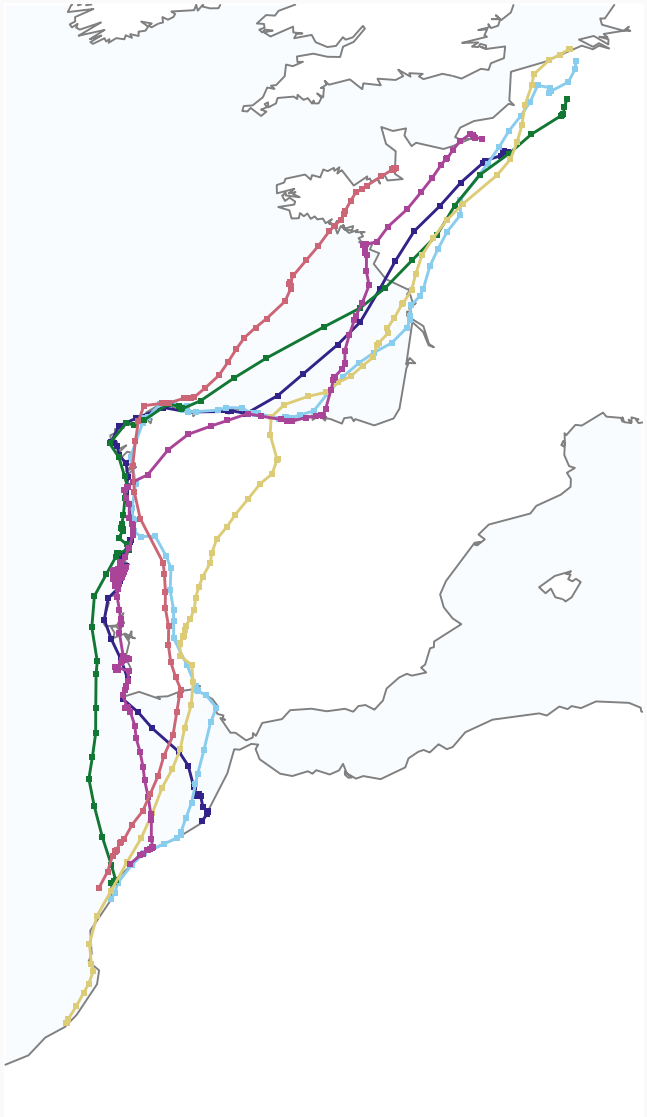


Image data

Image data

Image data


131 135 131 134 152 59 40 41 15 13 46 33 2 10 18 27 48 30 26 34 26 181 178 182 178
149 138 128 133 159 116 60 90 110 79 49 41 41 15 0 31 33 25 25 34 35 85 163 165 195
134 133 138 139 139 166 108 121 125 146 146 131 128 140 112 36 18 22 32 21 25 26 118 133 170
145 138 138 148 124 168 154 121 91 113 134 150 142 154 160 98 25 43 11 24 34 32 89 131 117
158 168 131 147 147 149 172 124 111 129 119 137 147 131 145 164 168 166 68 71 84 63 59 111 139
78 167 149 139 159 144 165 122 109 128 137 141 136 132 135 124 138 159 165 62 25 37 12 136 158
24 98 156 146 148 156 151 116 127 98 128 151 129 125 132 127 153 138 173 76 35 64 16 104 174
76 19 123 165 142 167 138 119 118 111 129 133 138 133 121 149 139 154 145 182 92 28 69 141 177
87 73 37 138 158 170 113 121 105 114 119 129 138 132 128 135 153 126 159 163 143 91 129 174 179
85 71 70 34 151 145 128 116 144 118 101 129 151 127 110 130 133 157 160 168 145 102 169 193 198
97 99 84 57 56 136 117 122 132 129 111 119 139 127 115 135 133 181 152 179 140 107 203 202 210
130 113 107 122 100 159 120 97 111 133 122 111 127 128 123 138 138 163 133 181 107 108 217 202 210
85 112 122 124 145 135 135 159 136 137 119 113 125 120 115 133 131 135 144 183 55 122 228 208 212
129 137 107 122 129 131 115 114 142 152 143 119 106 110 119 131 125 143 189 193 22 150 239 216 212
80 92 125 139 145 145 139 131 134 153 160 134 99 105 124 129 128 161 208 201 25 179 235 220 203
65 82 80 94 127 152 101 146 148 129 141 153 125 106 115 124 123 155 187 196 38 190 222 223 197
69 62 81 85 90 143 169 131 124 134 119 138 136 118 96 125 130 126 174 167 2 190 231 218 181
66 63 81 88 85 118 149 138 135 166 160 151 151 138 106 92 121 131 185 147 36 174 216 213 162
65 64 71 78 76 94 131 158 169 168 137 107 136 152 128 94 83 145 197 155 88 199 217 223 148
59 64 70 72 73 80 108 140 139 155 168 154 162 140 124 110 86 99 154 161 20 185 236 184 140
55 69 74 72 75 71 73 87 125 142 167 162 163 135 134 139 98 95 100 139 49 199 222 182 132
62 71 67 59 59 60 60 65 80 97 126 149 182 201 212 209 151 89 75 112 76 183 212 148 126
70 70 66 57 50 50 59 67 66 74 71 62 49 72 103 124 178 153 92 61 87 163 186 141 116
71 71 80 79 60 47 50 58 81 86 90 105 84 86 78 84 67 96 83 52 59 156 187 95 97
75 79 82 78 54 61 111 135 154 164 181 199 213 220 221 223 225 203 187 147 51 106 134 77 82

Image data

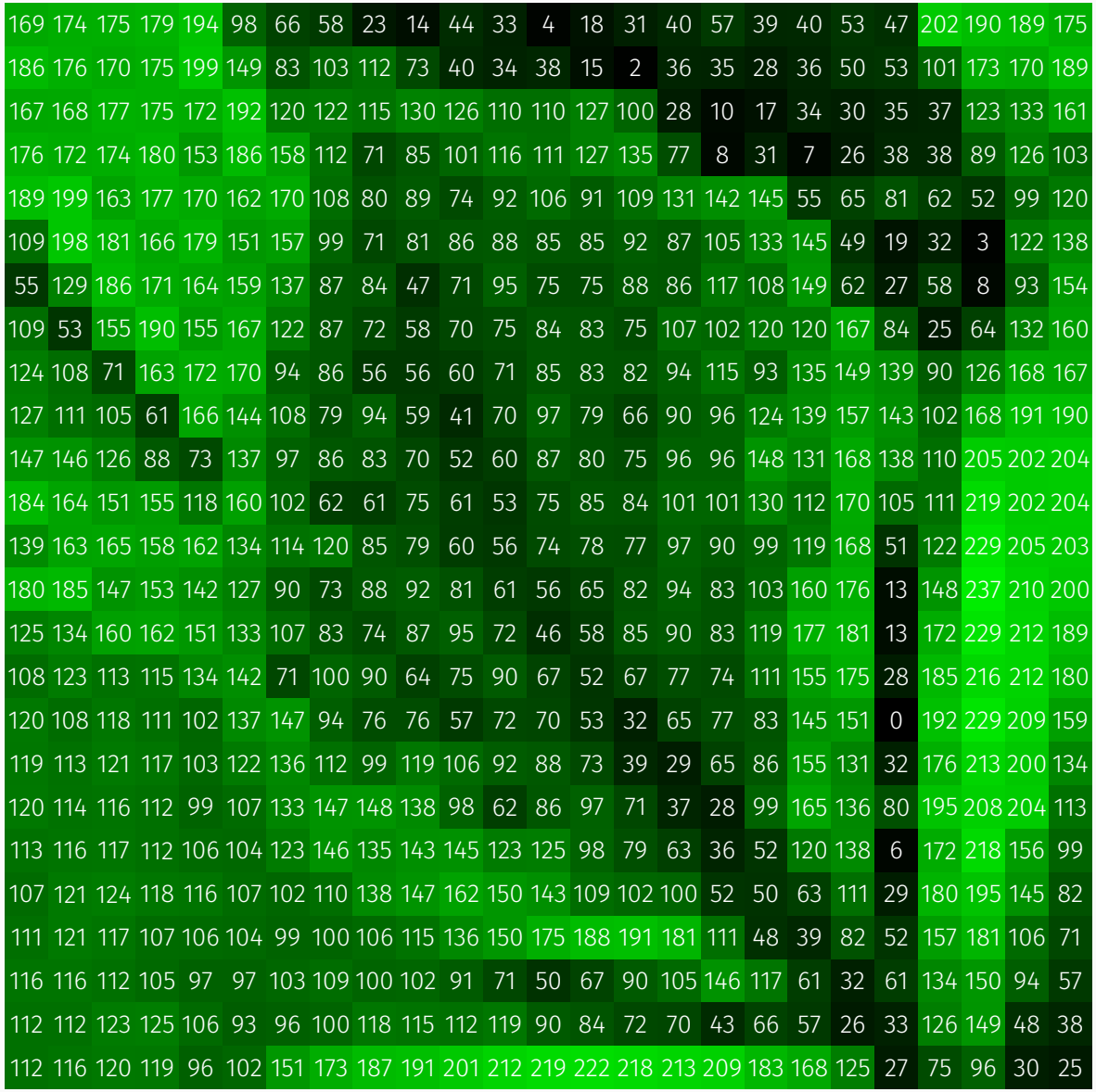
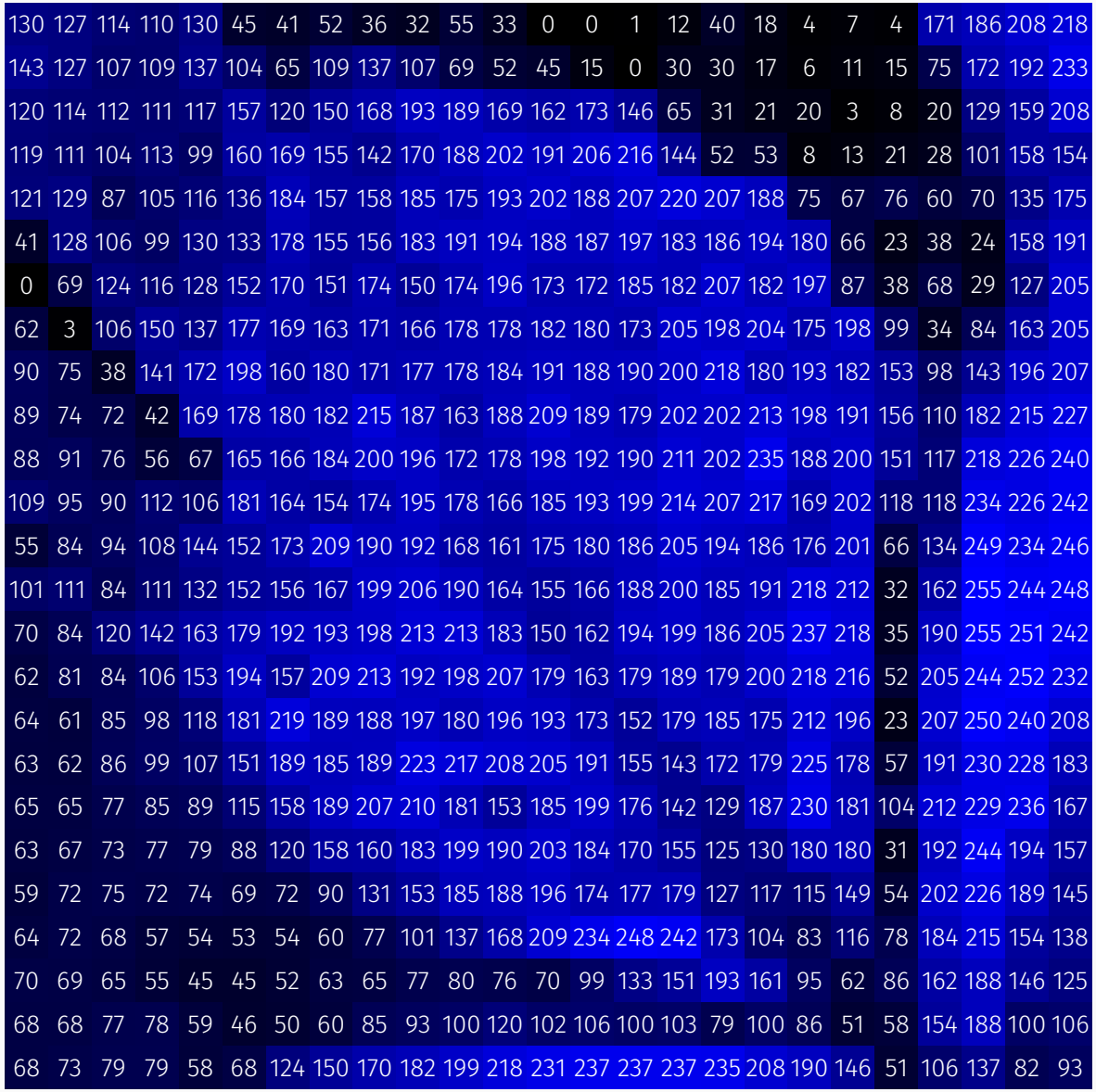

169 174 175 179 194 98 66 58 23 14 44 33 4 18 31 40 57 39 40 53 47 202 190 189 175
186 176 170 175 199 149 83 103 112 73 40 34 38 15 2 36 35 28 36 50 53 101 173 170 189
167 168 177 175 172 192 120 122 115 130 126 110 110 127 100 28 10 17 34 30 35 37 123 133 161
176 172 174 180 153 186 158 112 71 85 101 116 111 127 135 77 8 31 7 26 38 38 89 126 103
189 199 163 177 170 162 170 108 80 89 74 92 106 91 109 131 142 145 55 65 81 62 52 99 120
109 198 181 166 179 151 157 99 71 81 86 88 85 85 92 87 105 133 145 49 19 32 3 122 138
55 129 186 171 164 159 137 87 84 47 71 95 75 75 88 86 117 108 149 62 27 58 8 93 154
109 53 155 190 155 167 122 87 72 58 70 75 84 83 75 107 102 120 120 167 84 25 64 132 160
124 108 71 163 172 170 94 86 56 56 60 71 85 83 82 94 115 93 135 149 139 90 126 168 167
127 111 105 61 166 144 108 79 94 59 41 70 97 79 66 90 96 124 139 157 143 102 168 191 190
147 146 126 88 73 137 97 86 83 70 52 60 87 80 75 96 96 148 131 168 138 110 205 202 204
184 164 151 155 118 160 102 62 61 75 61 53 75 85 84 101 101 130 112 170 105 111 219 202 204
139 163 165 158 162 134 114 120 85 79 60 56 74 78 77 97 90 99 119 168 51 122 229 205 203
180 185 147 153 142 127 90 73 88 92 81 61 56 65 82 94 83 103 160 176 13 148 237 210 200
125 134 160 162 151 133 107 83 74 87 95 72 46 58 85 90 83 119 177 181 13 172 229 212 189
108 123 113 115 134 142 71 100 90 64 75 90 67 52 67 77 74 111 155 175 28 185 216 212 180
120 108 118 111 102 137 147 94 76 76 57 72 70 53 32 65 77 83 145 151 0 192 229 209 159
119 113 121 117 103 122 136 112 99 119 106 92 88 73 39 29 65 86 155 131 32 176 213 200 134
120 114 116 112 99 107 133 147 148 138 98 62 86 97 71 37 28 99 165 136 80 195 208 204 113
113 116 117 112 106 104 123 146 135 143 145 123 125 98 79 63 36 52 120 138 6 172 218 156 99
107 121 124 118 116 107 102 110 138 147 162 150 143 109 102 100 52 50 63 111 29 180 195 145 82
111 121 117 107 106 104 99 100 106 115 136 150 175 188 191 181 111 48 39 82 52 157 181 106 71
116 116 112 105 97 97 103 109 100 102 91 71 50 67 90 105 146 117 61 32 61 134 150 94 57
112 112 123 125 106 93 96 100 118 115 112 119 90 84 72 70 43 66 57 26 33 126 149 48 38
112 116 120 119 96 102 151 173 187 191 201 212 219 222 218 213 209 183 168 125 27 75 96 30 25

Image data


130 127 114 110 130 45 41 52 36 32 55 33 0 0 1 12 40 18 4 7 4 171 186 208 218
143 127 107 109 137 104 65 109 137 107 69 52 45 15 0 30 30 17 6 11 15 75 172 192 233
120 114 112 111 117 157 120 150 168 193 189 169 162 173 146 65 31 21 20 3 8 20 129 159 208
119 111 104 113 99 160 169 155 142 170 188 202 191 206 216 144 52 53 8 13 21 28 101 158 154
121 129 87 105 116 136 184 157 158 185 175 193 202 188 207 220 207 188 75 67 76 60 70 135 175
41 128 106 99 130 133 178 155 156 183 191 194 188 187 197 183 186 194 180 66 23 38 24 158 191
0 69 124 116 128 152 170 151 174 150 174 196 173 172 185 182 207 182 197 87 38 68 29 127 205
62 3 106 150 137 177 169 163 171 166 178 178 182 180 173 205 198 204 175 198 99 34 84 163 205
90 75 38 141 172 198 160 180 171 177 178 184 191 188 190 200 218 180 193 182 153 98 143 196 207
89 74 72 42 169 178 180 182 215 187 163 188 209 189 179 202 202 213 198 191 156 110 182 215 227
88 91 76 56 67 165 166 184 200 196 172 178 198 192 190 211 202 235 188 200 151 117 218 226 240
109 95 90 112 106 181 164 154 174 195 178 166 185 193 199 214 207 217 169 202 118 118 234 226 242
55 84 94 108 144 152 173 209 190 192 168 161 175 180 186 205 194 186 176 201 66 134 249 234 246
101 111 84 111 132 152 156 167 199 206 190 164 155 166 188 200 185 191 218 212 32 162 255 244 248
70 84 120 142 163 179 192 193 198 213 213 183 150 162 194 199 186 205 237 218 35 190 255 251 242
62 81 84 106 153 194 157 209 213 192 198 207 179 163 179 189 179 200 218 216 52 205 244 252 232
64 61 85 98 118 181 219 189 188 197 180 196 193 173 152 179 185 175 212 196 23 207 250 240 208
63 62 86 99 107 151 189 185 189 223 217 208 205 191 155 143 172 179 225 178 57 191 230 228 183
65 65 77 85 89 115 158 189 207 210 181 153 185 199 176 142 129 187 230 181 104 212 229 236 167
63 67 73 77 79 88 120 158 160 183 199 190 203 184 170 155 125 130 180 180 31 192 244 194 157
59 72 75 72 74 69 72 90 131 153 185 188 196 174 177 179 127 117 115 149 54 202 226 189 145
64 72 68 57 54 53 54 60 77 101 137 168 209 234 248 242 173 104 83 116 78 184 215 154 138
70 69 65 55 45 45 52 63 65 77 80 76 70 99 133 151 193 161 95 62 86 162 188 146 125
68 68 77 78 59 46 50 60 85 93 100 120 102 106 100 103 79 100 86 51 58 154 188 100 106
68 73 79 79 58 68 124 150 170 182 199 218 231 237 237 237 235 208 190 146 51 106 137 82 93

Spatial data

grid data (only the order matters) *image*

geo-located data (explicit location) *demographic records*

regularly sampled *magnetic resonance imaging (MRI), positron emission tomography (PET)*

irregularly sampled *disease outbreaks, forest fires*

real values *surface temperature*

symbolic values *landcover records*

spatial *image, topographic records*

spatio-temporal *video, surface temperature, GPS traces*

Spatial data

grid vs. geo-located data

regularly vs. irregularly sampled

real vs. symbolic values

spatial vs. spatio-temporal

remote sensing, object tracking, geo-located services...

Spatial data

Spatial data can be viewed as contextual data

Contextual attribute(s) provide context for the measurements, reference points e.g. *geographic coordinates, incremental identifiers*

Behavioral attribute(s) represent the actual measurements

The dataset consists of n data points

$$\mathcal{D} = \langle (\boldsymbol{p}^{(1)}, \boldsymbol{x}^{(1)}), (\boldsymbol{p}^{(2)}, \boldsymbol{x}^{(2)}), \ldots, (\boldsymbol{p}^{(n)}, \boldsymbol{x}^{(n)}) \rangle$$

where $\boldsymbol{x}^{(i)} = \langle x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)} \rangle$ and $\boldsymbol{p}^{(i)} = \langle p_1^{(i)}, p_2^{(i)}, \ldots, p_c^{(i)} \rangle$ contain the values of the m behavioral attributes and of the c contextual attributes, respectively, for the i^{th} data point

Spatio-temporal data

Spatio-temporal data can be viewed as contextual data
Contextual attributes provide context for the measurements, reference points e.g. *date and geographic coordinates, incremental identifiers*
Behavioral attribute(s) represent the actual measurements

The dataset consists of n data points

$$\mathcal{D} = \langle (\boldsymbol{p}^{(1)}, \boldsymbol{x}^{(1)}), (\boldsymbol{p}^{(2)}, \boldsymbol{x}^{(2)}), \ldots, (\boldsymbol{p}^{(n)}, \boldsymbol{x}^{(n)}) \rangle$$

where $\boldsymbol{x}^{(i)} = \langle x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)} \rangle$ and $\boldsymbol{p}^{(i)} = \langle p_1^{(i)}, p_2^{(i)}, \ldots, p_c^{(i)} \rangle$ contain the values of the m behavioral attributes and of the c contextual attributes, respectively, for the i^{th} data point

Spatio-temporal data

The dataset consists of n data points

$$\mathcal{D} = \langle (\boldsymbol{p}^{(1)}, \boldsymbol{x}^{(1)}), (\boldsymbol{p}^{(2)}, \boldsymbol{x}^{(2)}), \ldots, (\boldsymbol{p}^{(n)}, \boldsymbol{x}^{(n)}) \rangle$$

where $\boldsymbol{x}^{(i)} = \langle x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)} \rangle$ and $\boldsymbol{p}^{(i)} = \langle p_1^{(i)}, p_2^{(i)}, \ldots, p_c^{(i)} \rangle$ contain the values of the m behavioral attributes and of the c contextual attributes, respectively, for the i^{th} data point

Regularly sampled univariate data with two-dimensional coordinates can be represented as a two-dimensional array, i.e. matrix, and displayed as an image

Regularly sampled univariate data with c-dimensional coordinates can be represented as a c-dimensional array, i.e. tensor

Measuring distances

The distance between the locations of two data points i and j, $\mathrm{d}(p^{(i)}, p^{(j)})$, might be measured using e.g. Euclidean or Manhattan distance

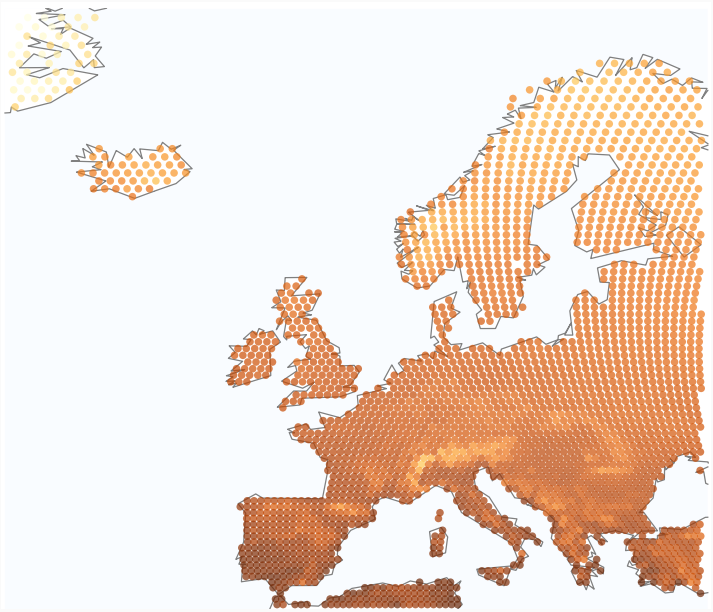
Coordinates might be provided as latitude and longitude

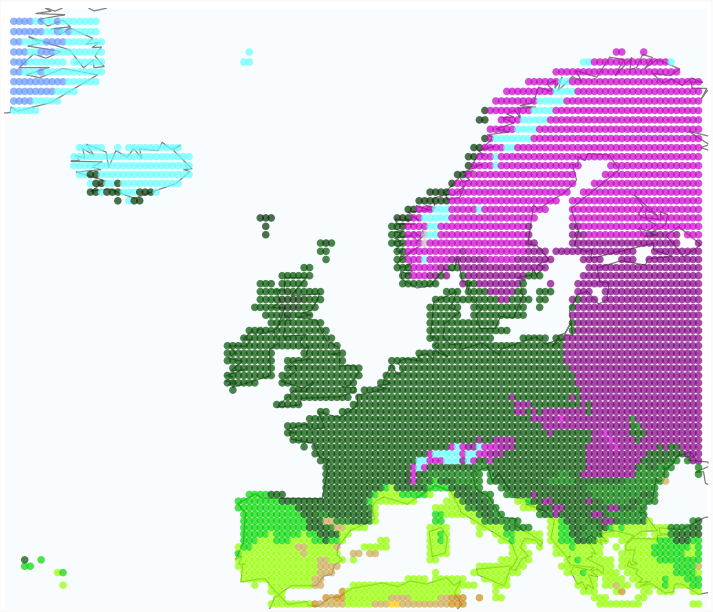
! The length of a degree of longitude varies with the latitude

Distances might be best measured using the *great circle distance* (a.k.a. orthodromic distance)

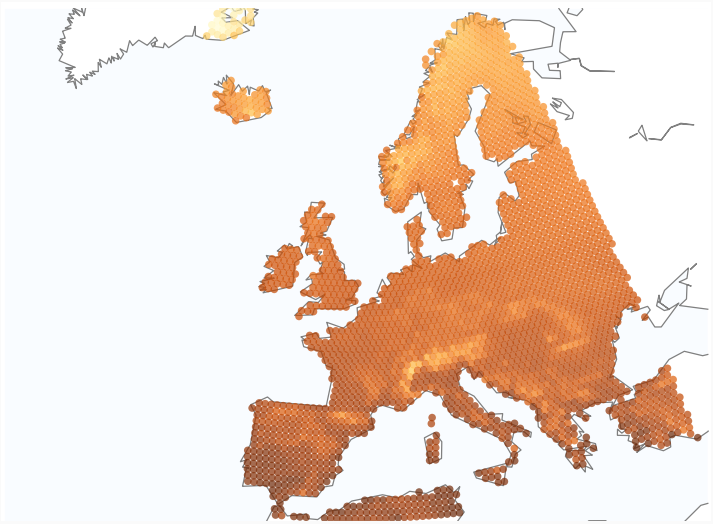
Distances, angles, surfaces and map projections

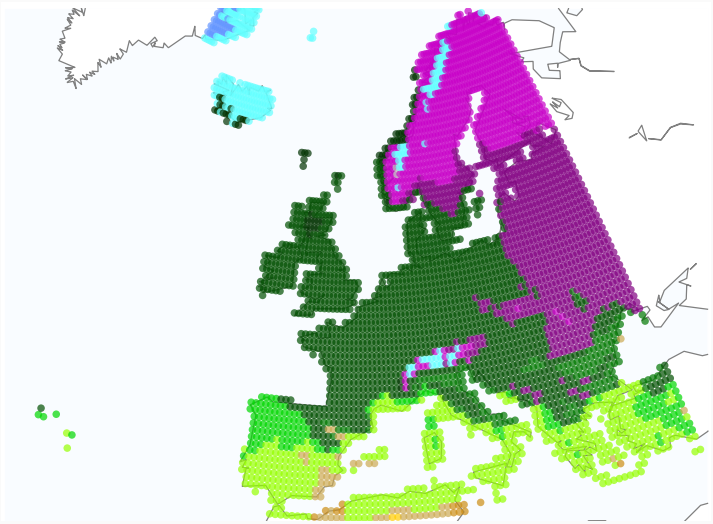
Miller





Stereographic





Distances, angles, surfaces and map projections

Different map projections have different properties

Conformal preserves angles and scale locally

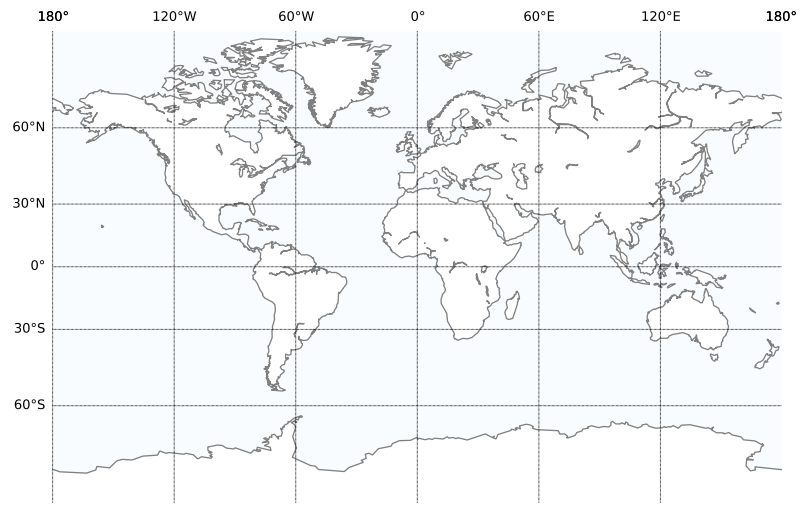
Equivalent preserves areas globally

Equidistant preserves all distances from one (or two) points

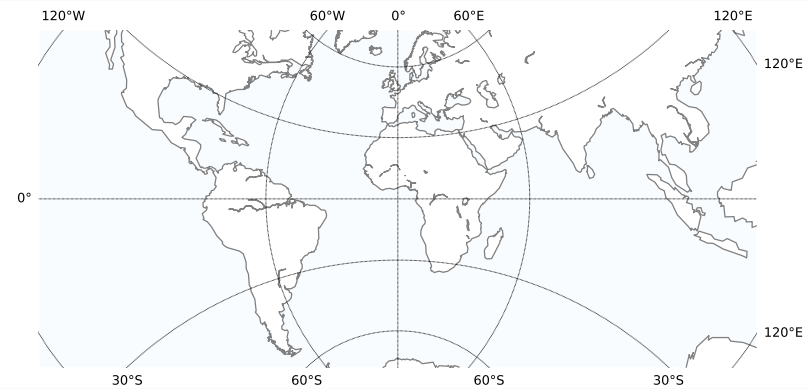
Compromise neither conformal nor equivalent,
aims to reduce overall distortion

Distances, angles, surfaces and map projections

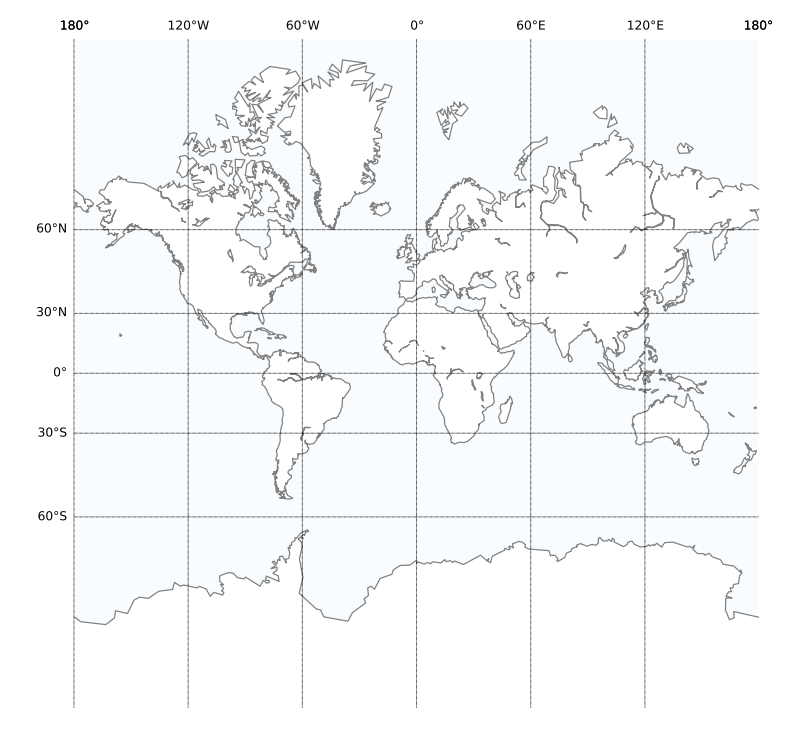
Miller (compromise)



Stereographic (conformal)

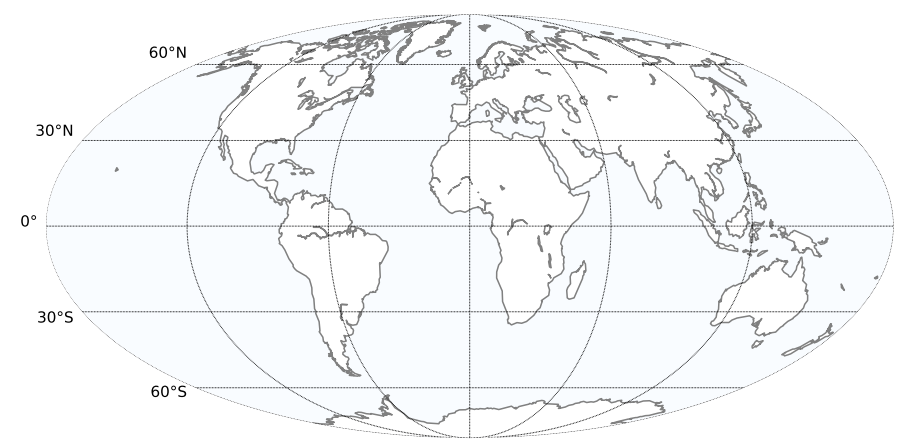


Mercator (conformal)

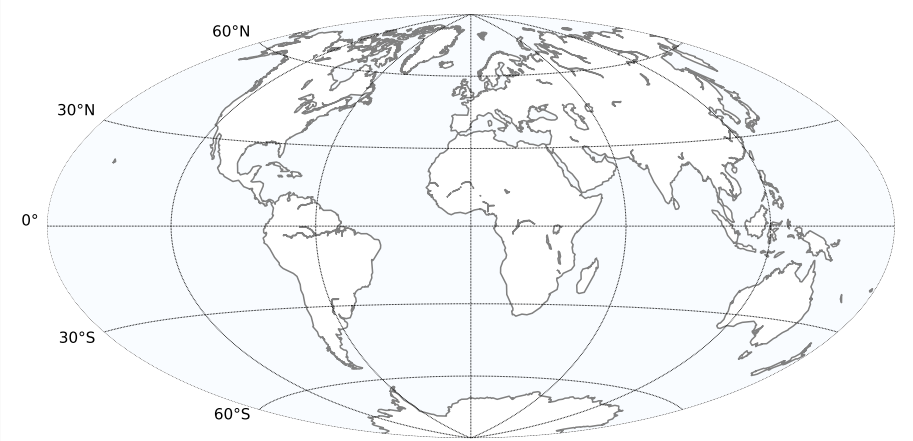


Distances, angles, surfaces and map projections

Mollweide (equivalent)

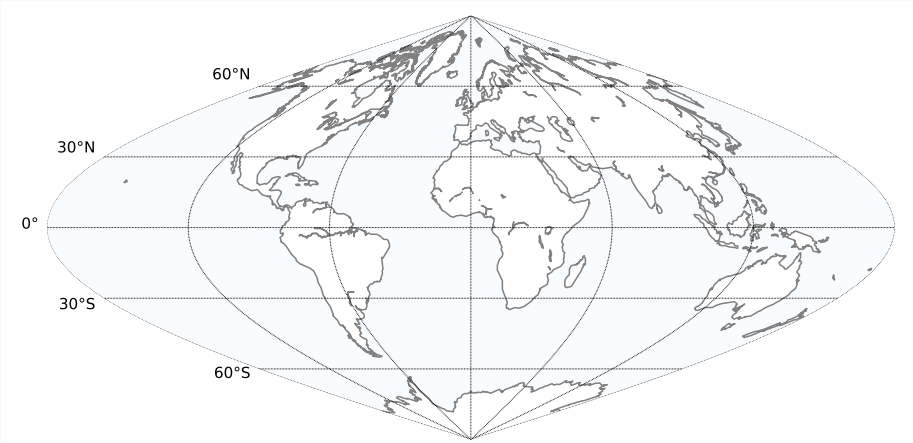


Hammer (equivalent)



Robinson (compromise)

Sinusoidal (equivalent, equidistant)



Interpolation

Interpolation can be used to produce a dataset with equally spaced coordinates, i.e. arranged along a grid
Map datasets from different grids, e.g. with different resolutions, to common grid

Inverse distance weighting

Let v_p denote the value at the point with coordinates p
Given a sample of point coordinates P for which the values are known, the value at coordinates q is estimated as

$$v_q = \begin{cases} v_p & \text{if } \exists\, p \in P, \mathrm{d}(q,p) = 0 \\ \dfrac{\sum_{p \in P} v_p / \mathrm{d}(q,p)}{\sum_{p \in P} 1 / \mathrm{d}(q,p)} & \text{otherwise} \end{cases}$$

Density estimation

Considering discrete attribute j and a value a in its domain, we collect in P the coordinates of data points that are occurrences of the corresponding item, i.e.

$$P = \{\boldsymbol{p}^{(i)} \text{ for } i = 1 \ldots n, \text{ such that } \boldsymbol{x}_j^{(i)} = a\}$$

Kernel density estimation methods produce density profiles, similarly to histogram techniques, but applying smoothing

The density of the item at coordinates $\boldsymbol{q}$ is estimated as

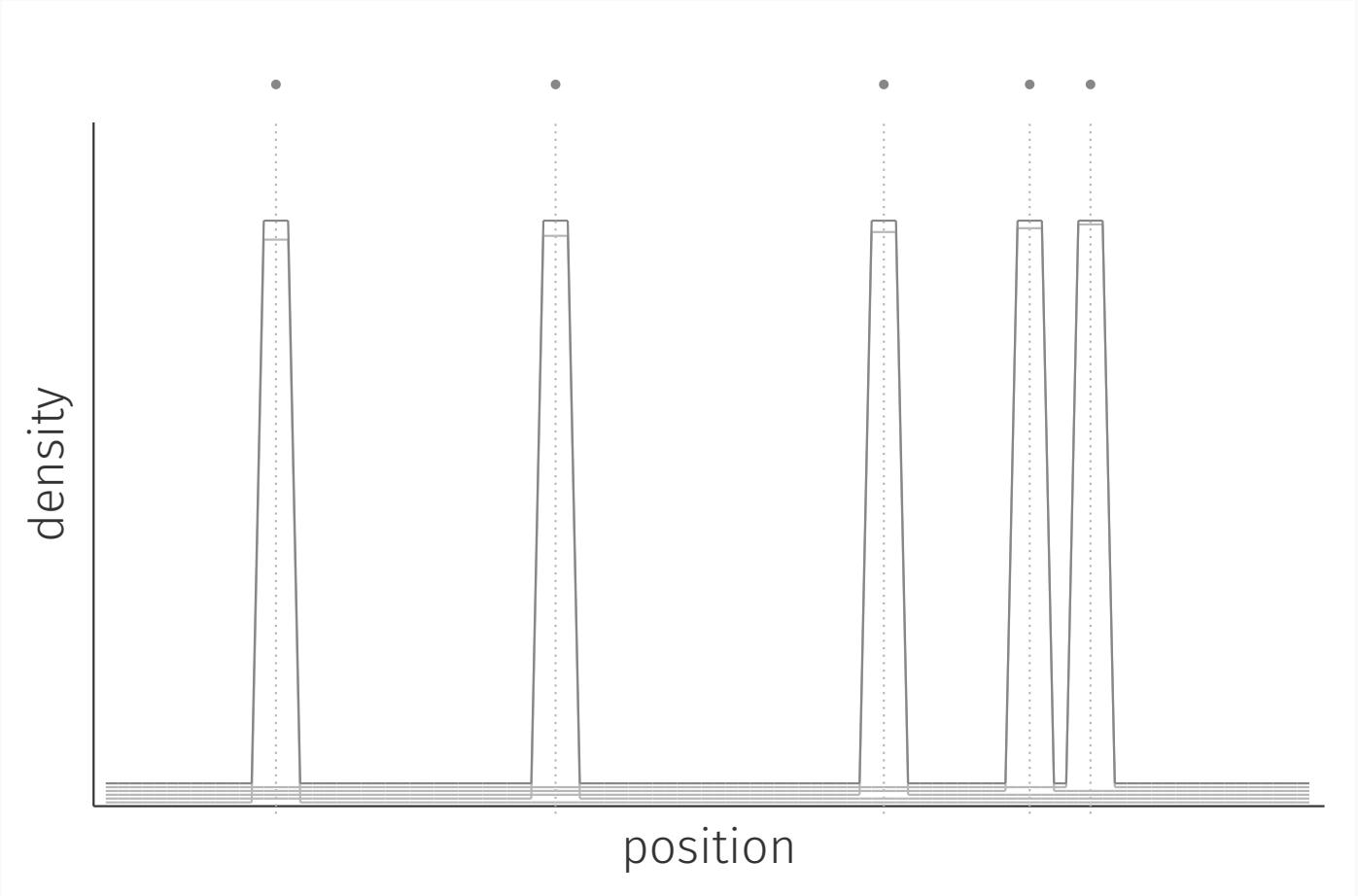
$$v_q = \frac{1}{|P|} \sum_{\boldsymbol{p} \in P} K_h(\boldsymbol{q}, \boldsymbol{p})$$

using for instance the Gaussian kernel of width h

$$K_h(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{(\sqrt{2\pi} \cdot h)^c} e^{-\|\boldsymbol{q}-\boldsymbol{p}\|_2^2/(2h^2)}$$

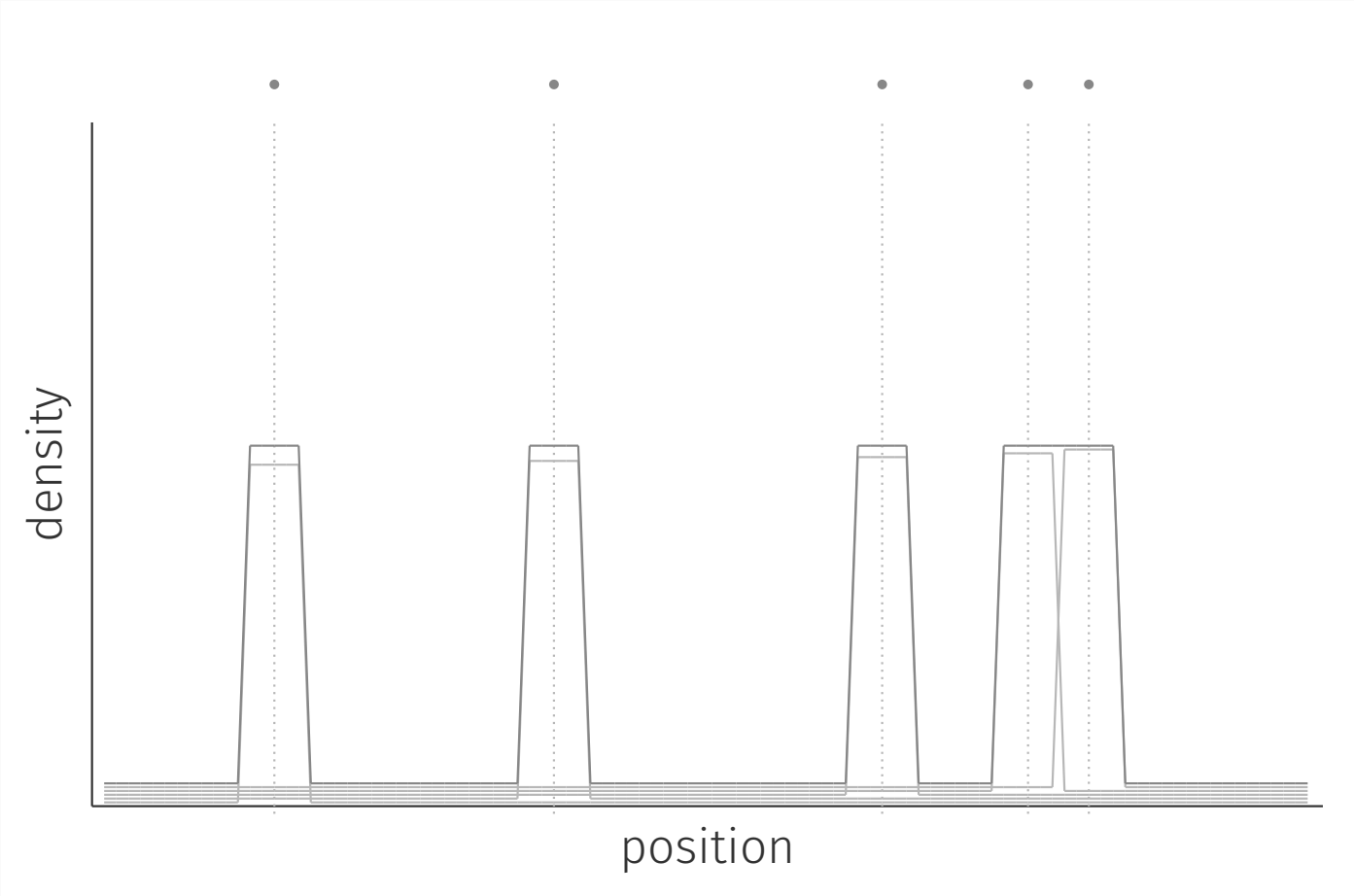
Density estimation

with a uniform window of width 3



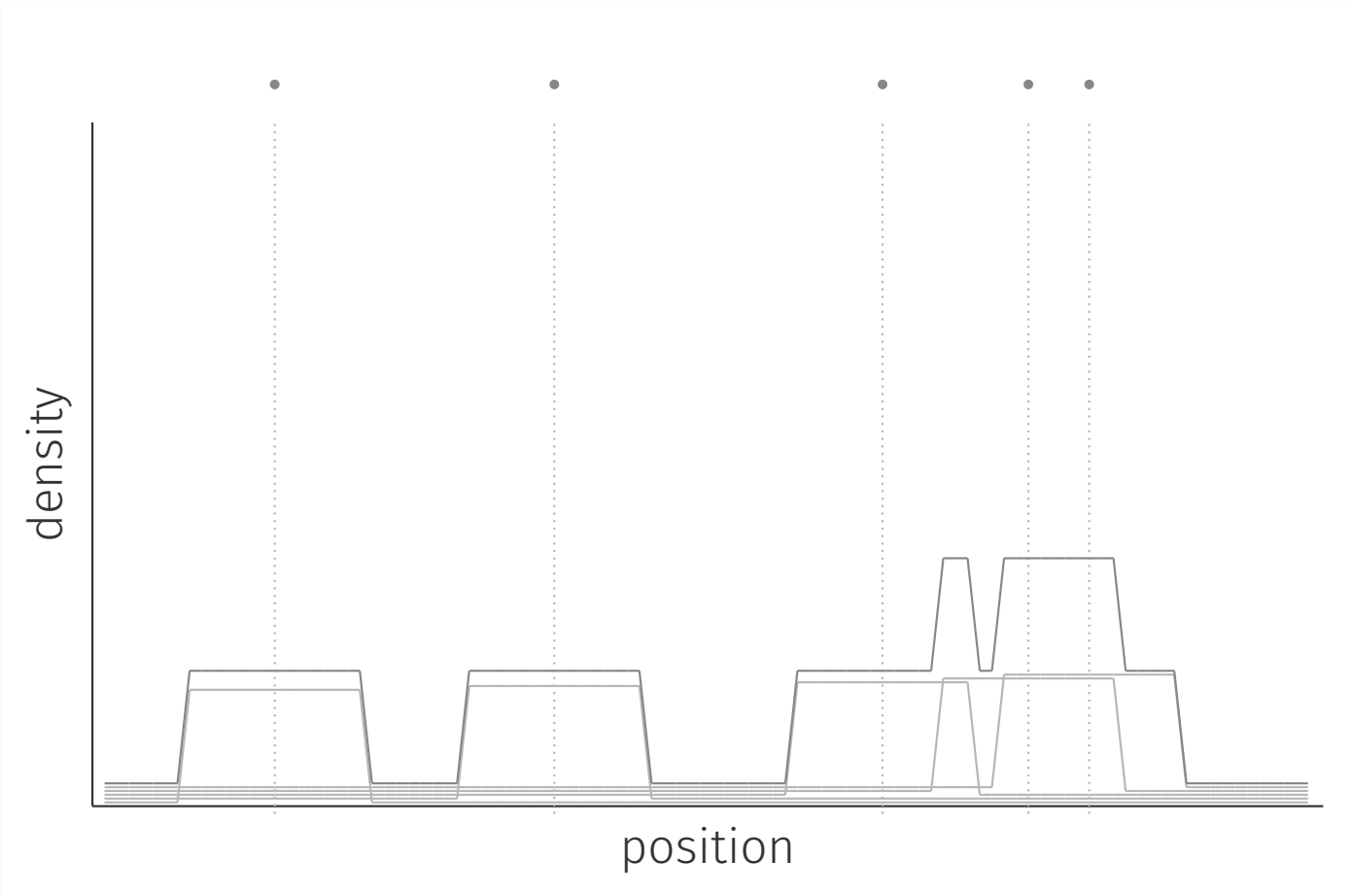
Density estimation

with a uniform window of width 5



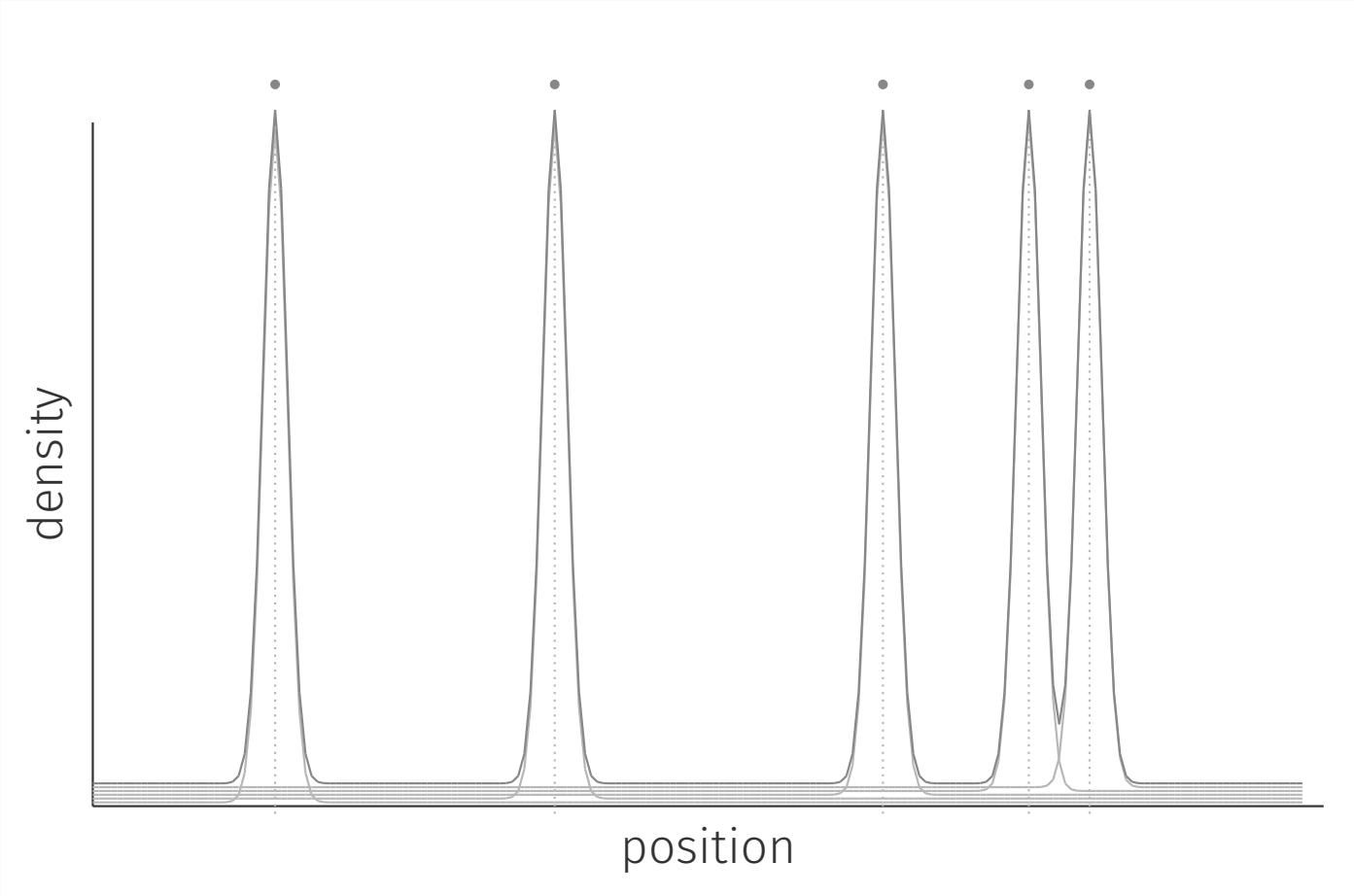
Density estimation

with a uniform window of width 15



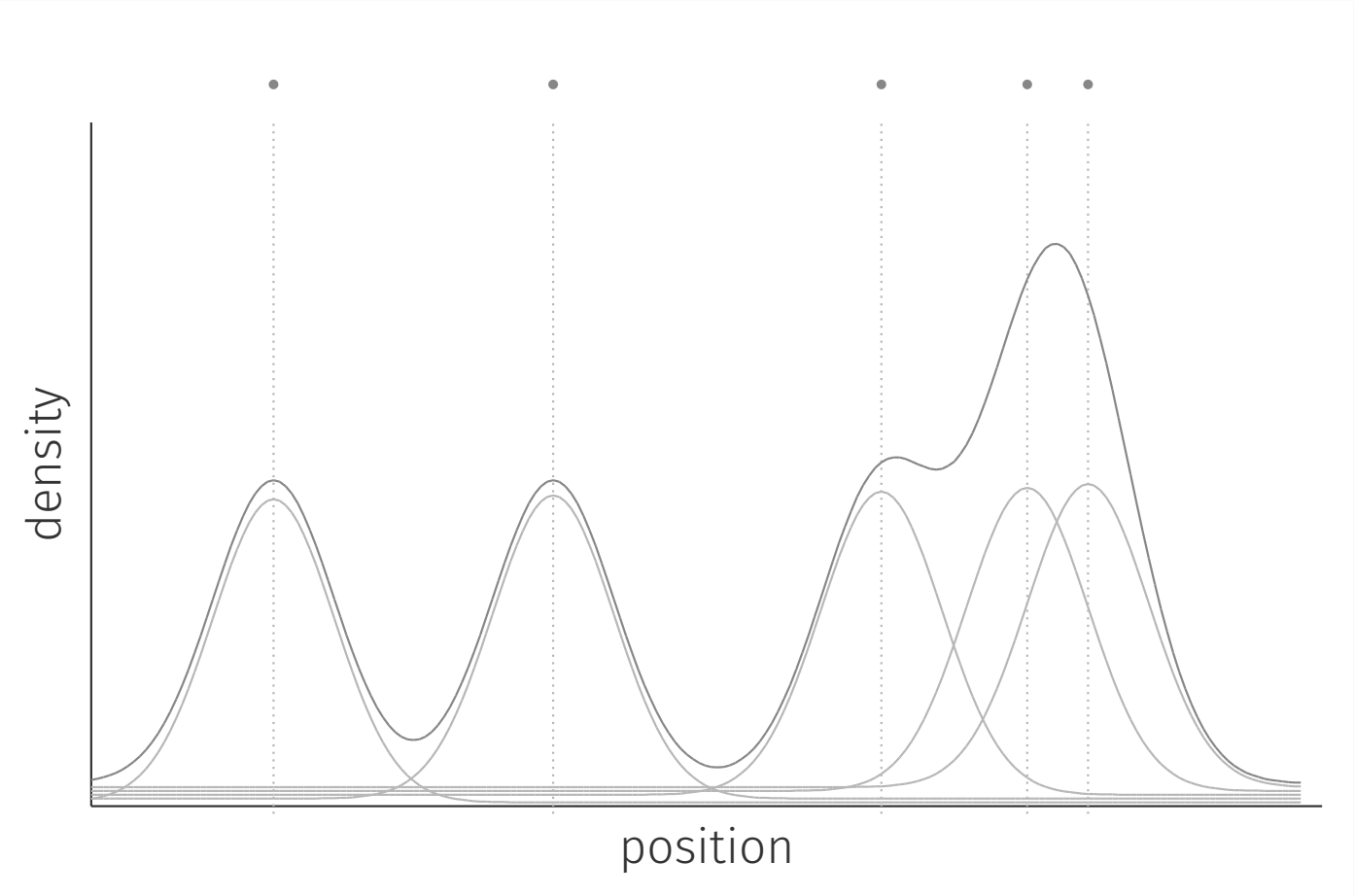
Density estimation

with a Gaussian kernel of width 1



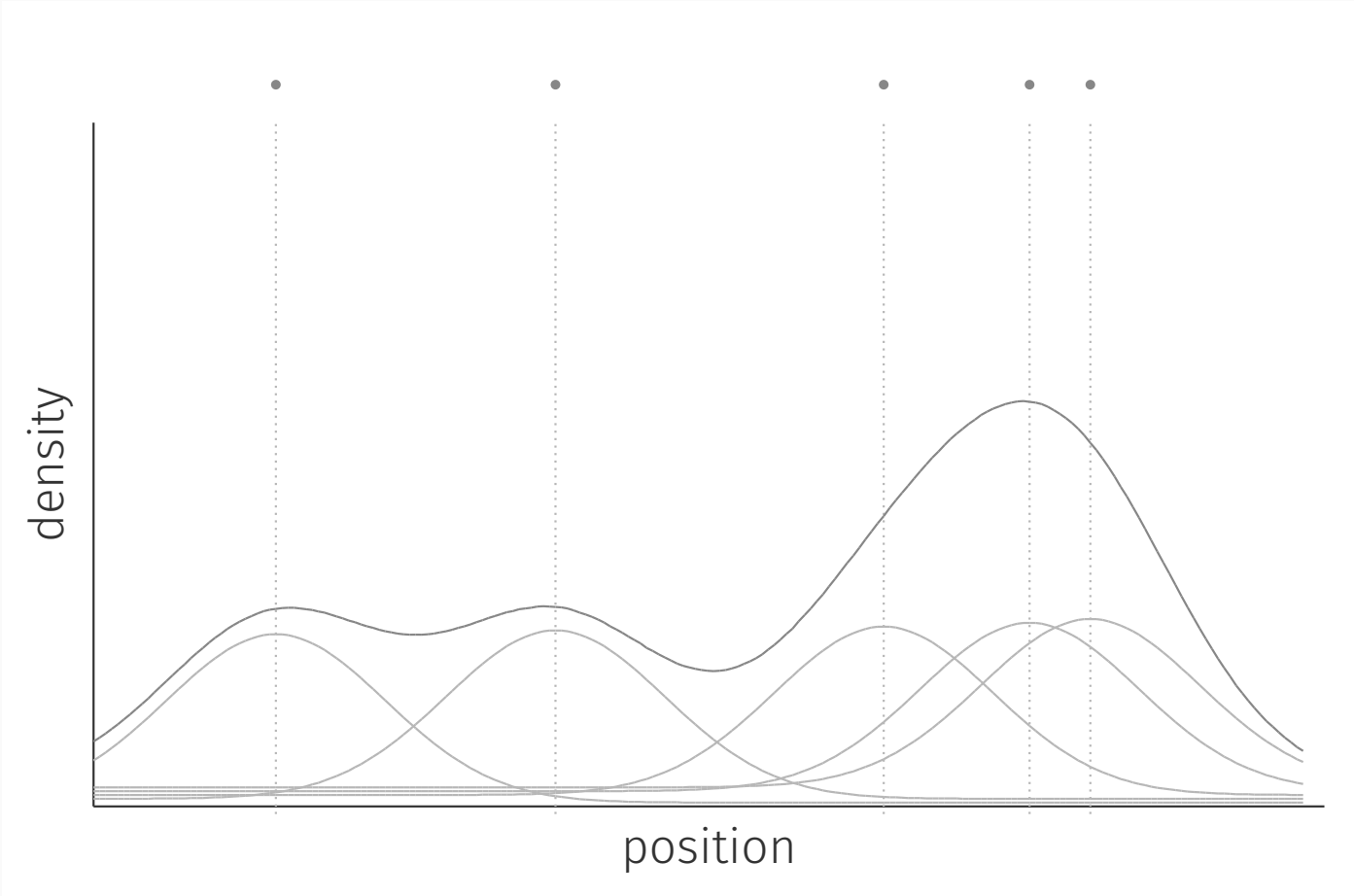
Density estimation

with a Gaussian kernel of width 5



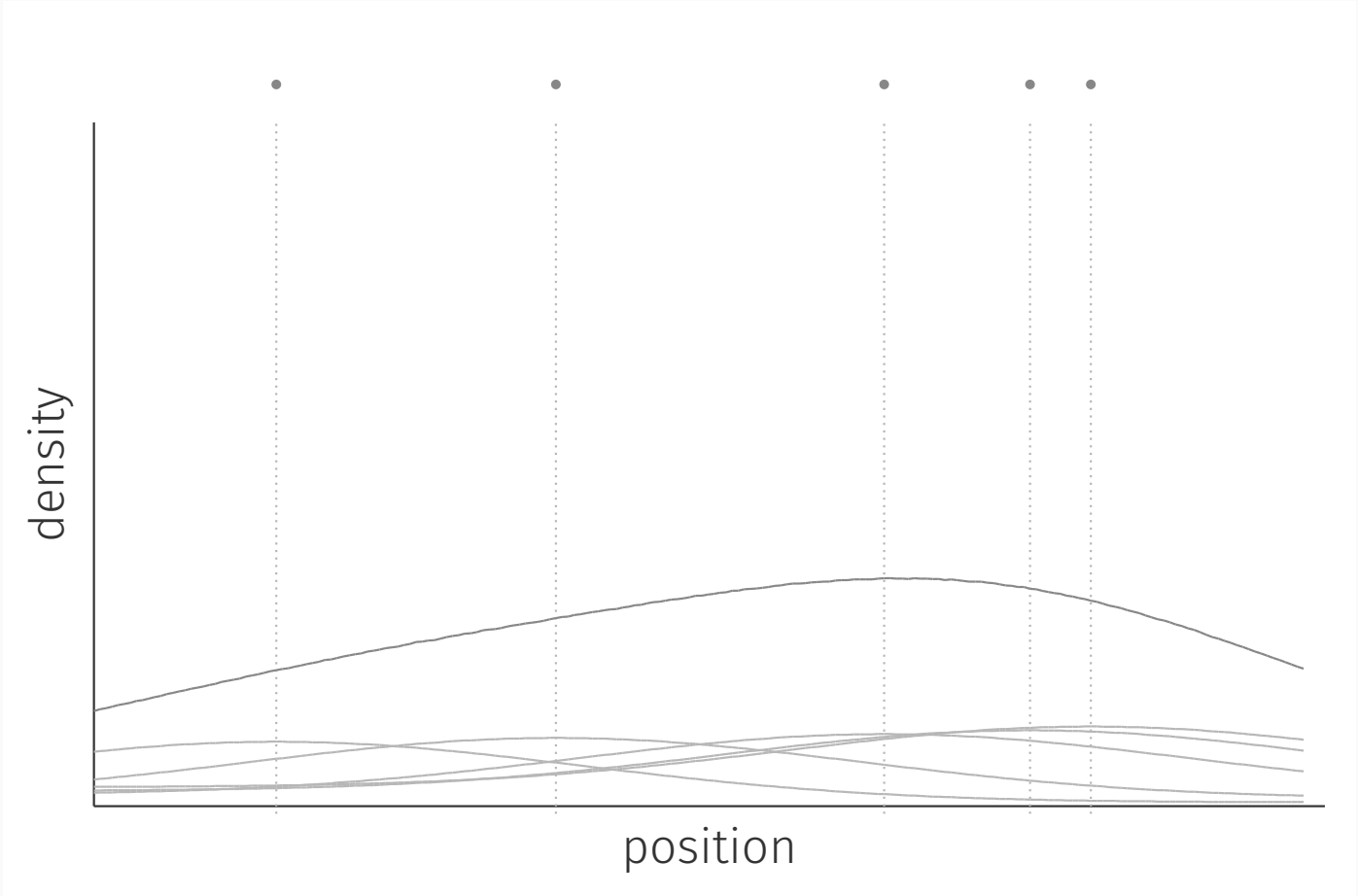
Density estimation

with a Gaussian kernel of width 9

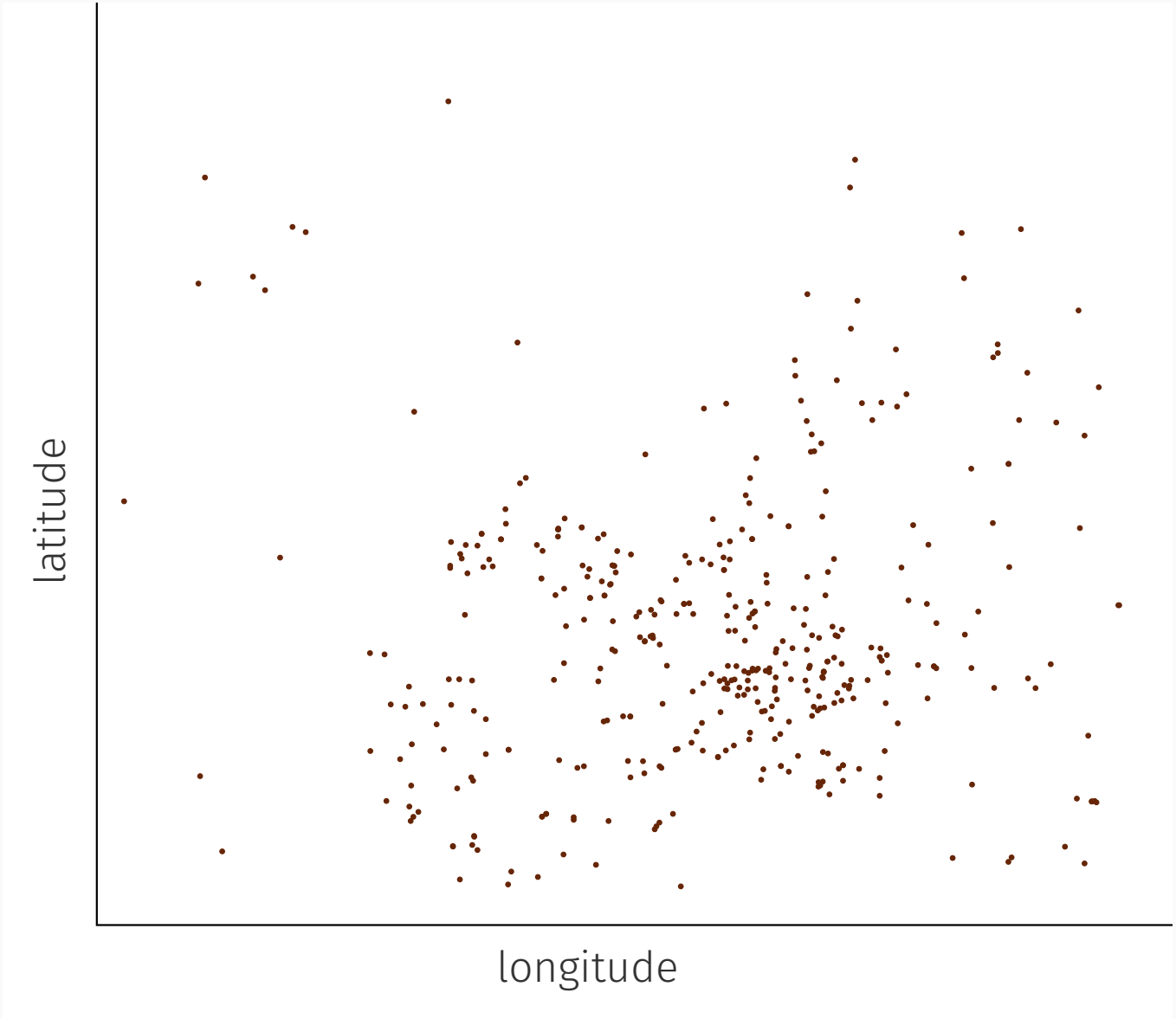


Density estimation

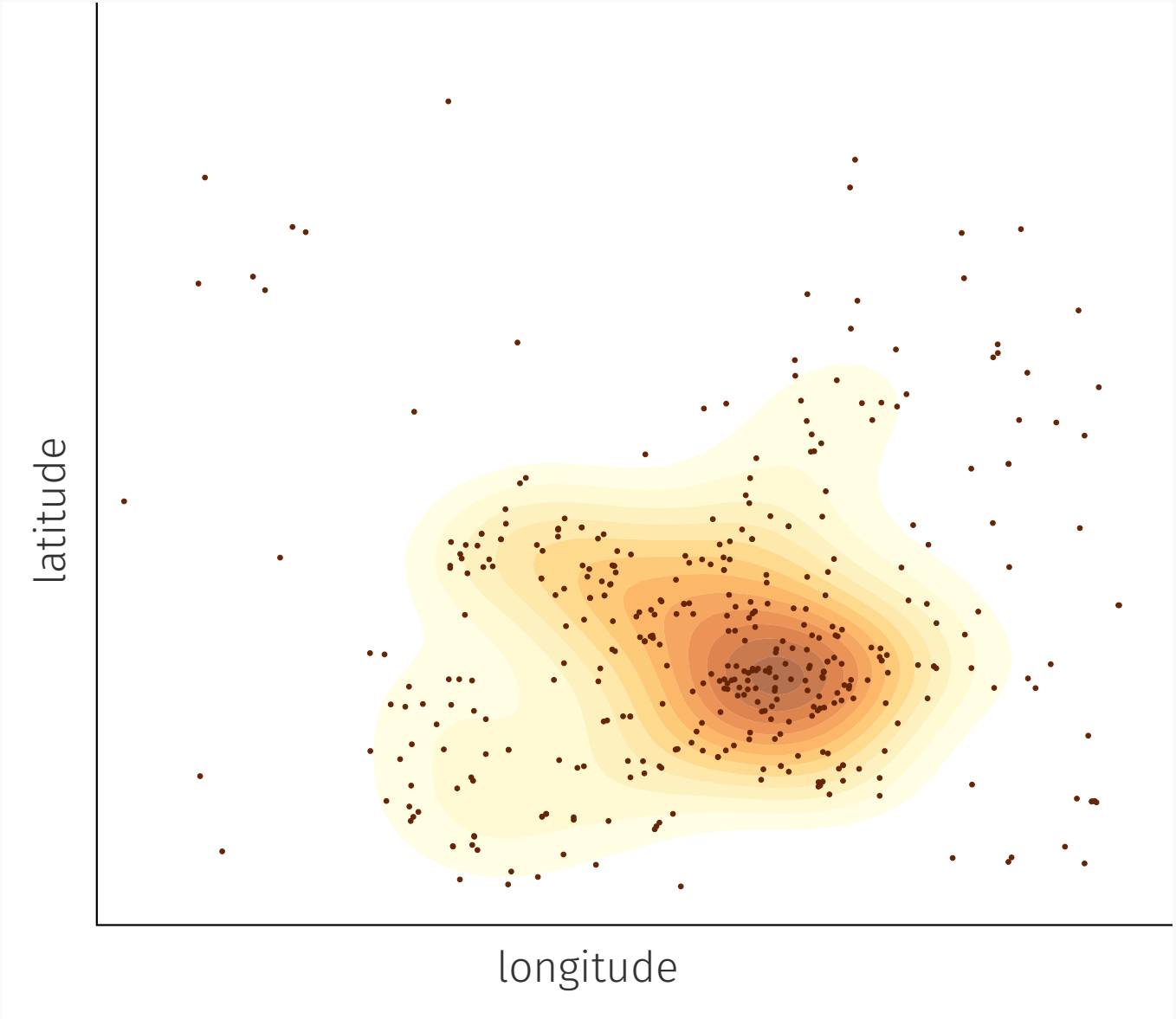
with a Gaussian kernel of width 25



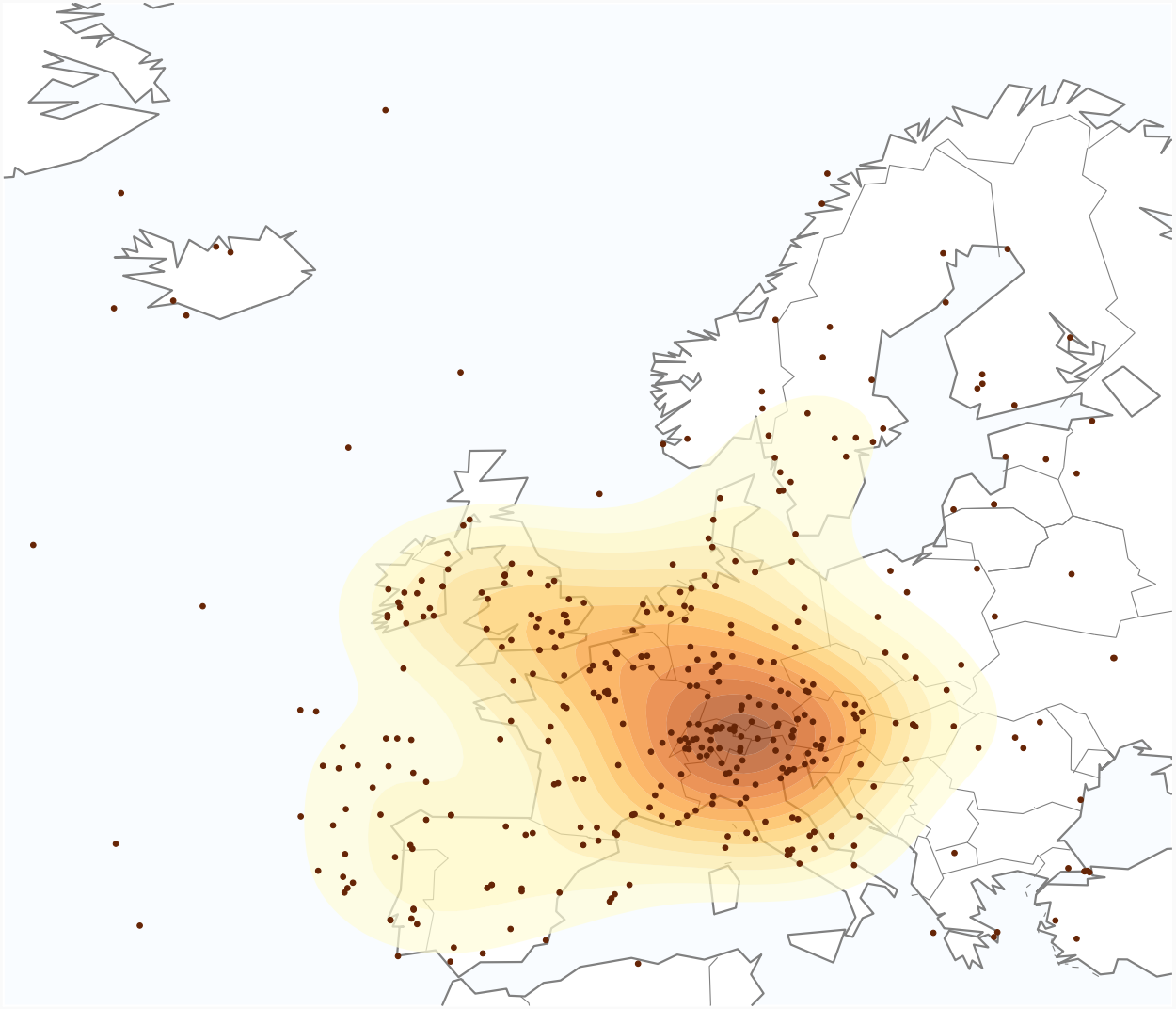
Aviation accidents



Aviation accidents: kernel density estimation



Aviation accidents: kernel density estimation



Triangulation

A *triangulation* divides the plane into triangles

The **Delaunay triangulation** for a set of points P is a triangulation $\mathcal{T}$ such that no point in P is inside the circumcircle of any triangle in $\mathcal{T}$

It maximizes the minimum angle among the triangles in $\mathcal{T}$
The *Euclidean minimum spanning tree* of a set of points is a subset of its Delaunay triangulation

Triangulation

A *triangulation* divides the plane into triangles

The **Delaunay triangulation** for a set of points P is a triangulation $\mathcal{T}$ such that no point in P is inside the circumcircle of any triangle in $\mathcal{T}$

It maximizes the minimum angle among the triangles in $\mathcal{T}$
The *Euclidean minimum spanning tree* of a set of points is a subset of its Delaunay triangulation

Algorithms to compute the Delaunay triangulation for a given set of points follow e.g. a divide-and-conquer approach ($O(n \log n)$) or an incremental approach ($O(n^2)$)

Triangulation

A *triangulation* divides the plane into triangles

The **Delaunay triangulation** for a set of points P is a triangulation $\mathcal{T}$ such that no point in P is inside the circumcircle of any triangle in $\mathcal{T}$

The **Voronoi diagram** of a set of points P is the dual of its Delaunay triangulation
It partitions the space into a collection of regions
Each point $\boldsymbol{p}$ in P is associated to a region (convex polygon) consisting of all points in the plane closer to $\boldsymbol{p}$ than to any other point of P

Triangulation

A *triangulation* divides the plane into triangles
The generalization of a triangle into higher-dimensional spaces is called *simplex*
A *triangulation* is a subdivision of the space into simplices

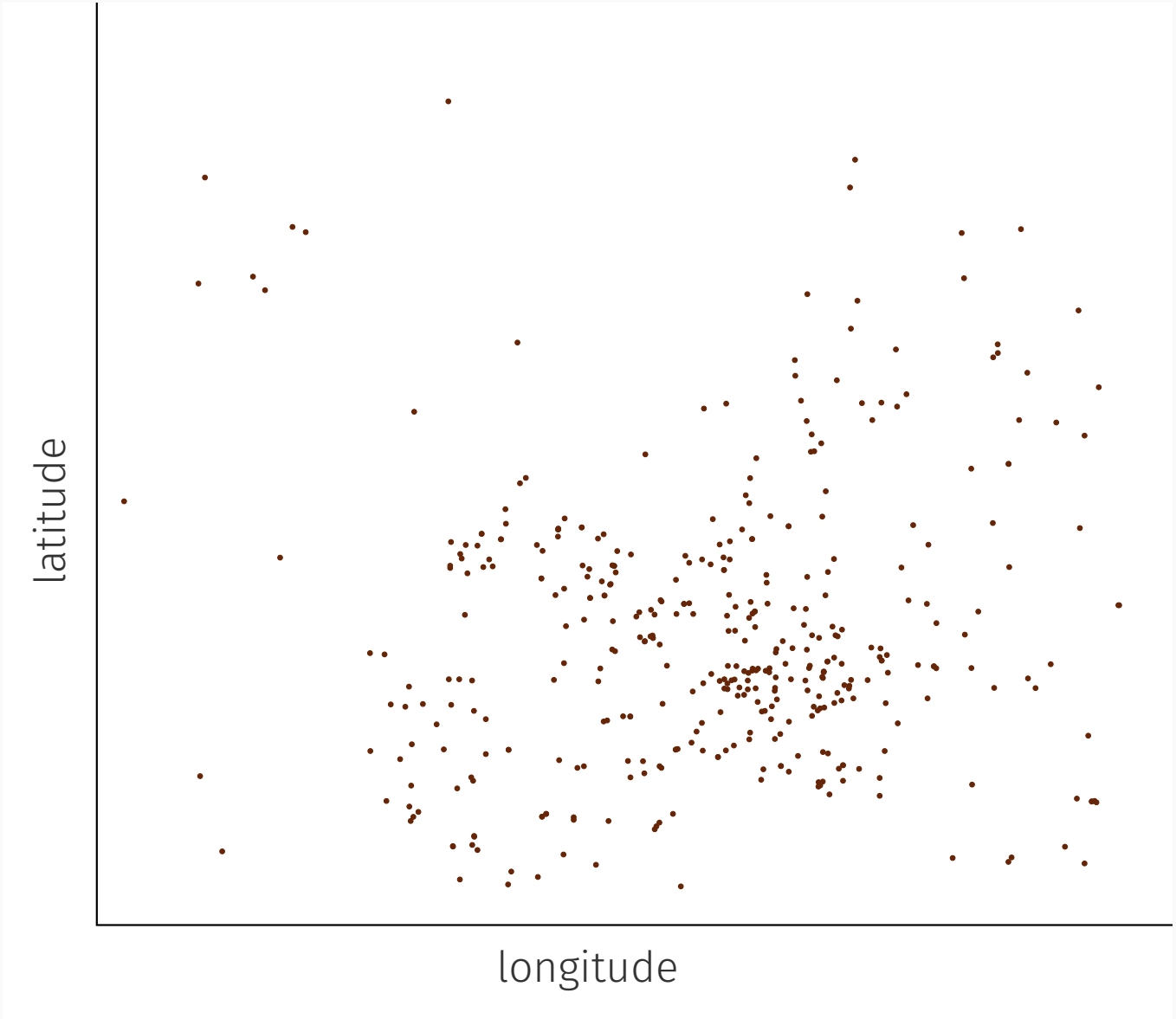
Delaunay triangulation and **Voronoi diagram** generalize to higher-dimensional spaces

Triangulation

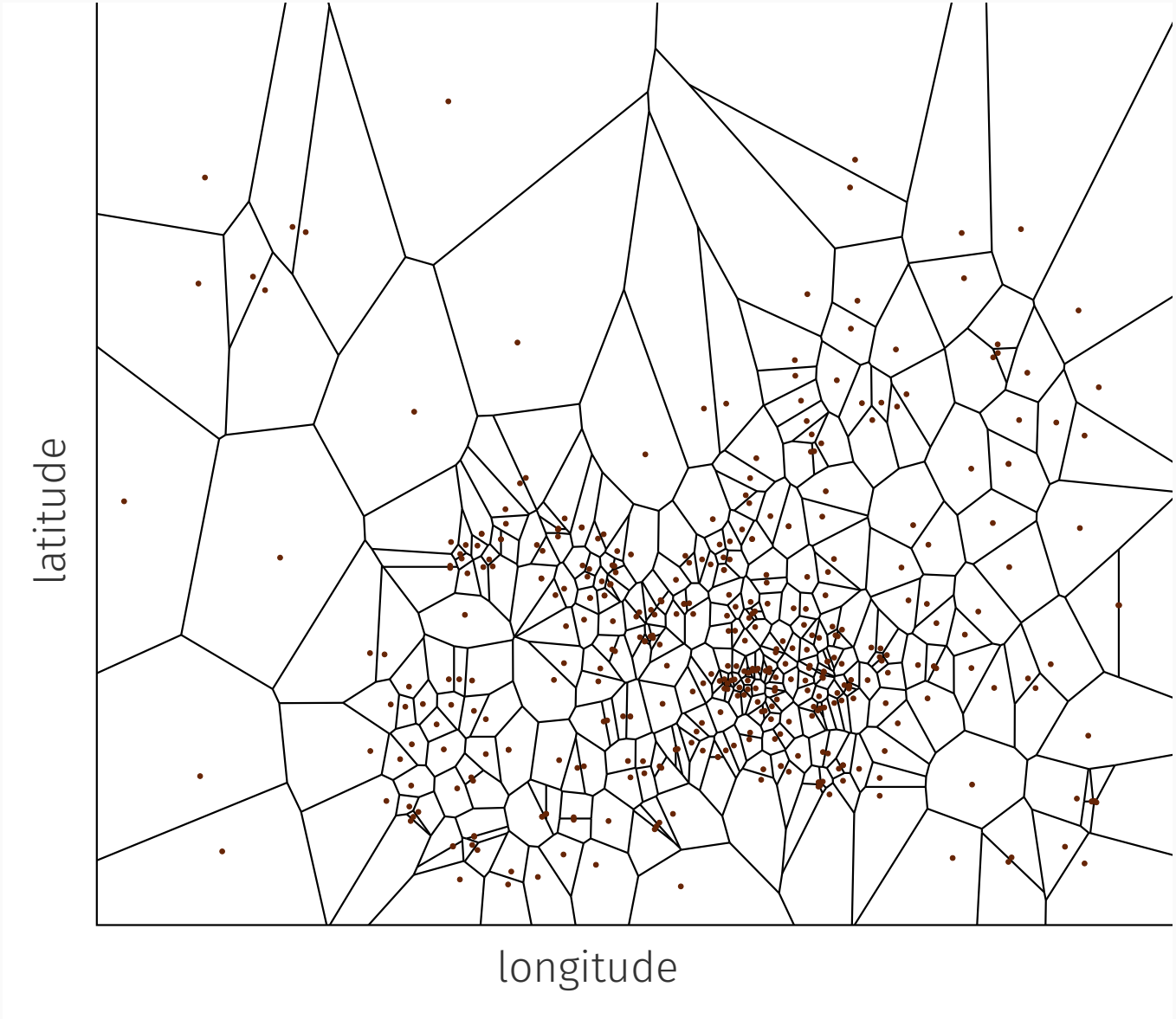
The **Delaunay triangulation** and the **Voronoi diagram** of a set of points P can be used to find the neighbors of a point, compute interpolated values, turn the data into a graph, etc.

They have multiple applications in a wide range of fields

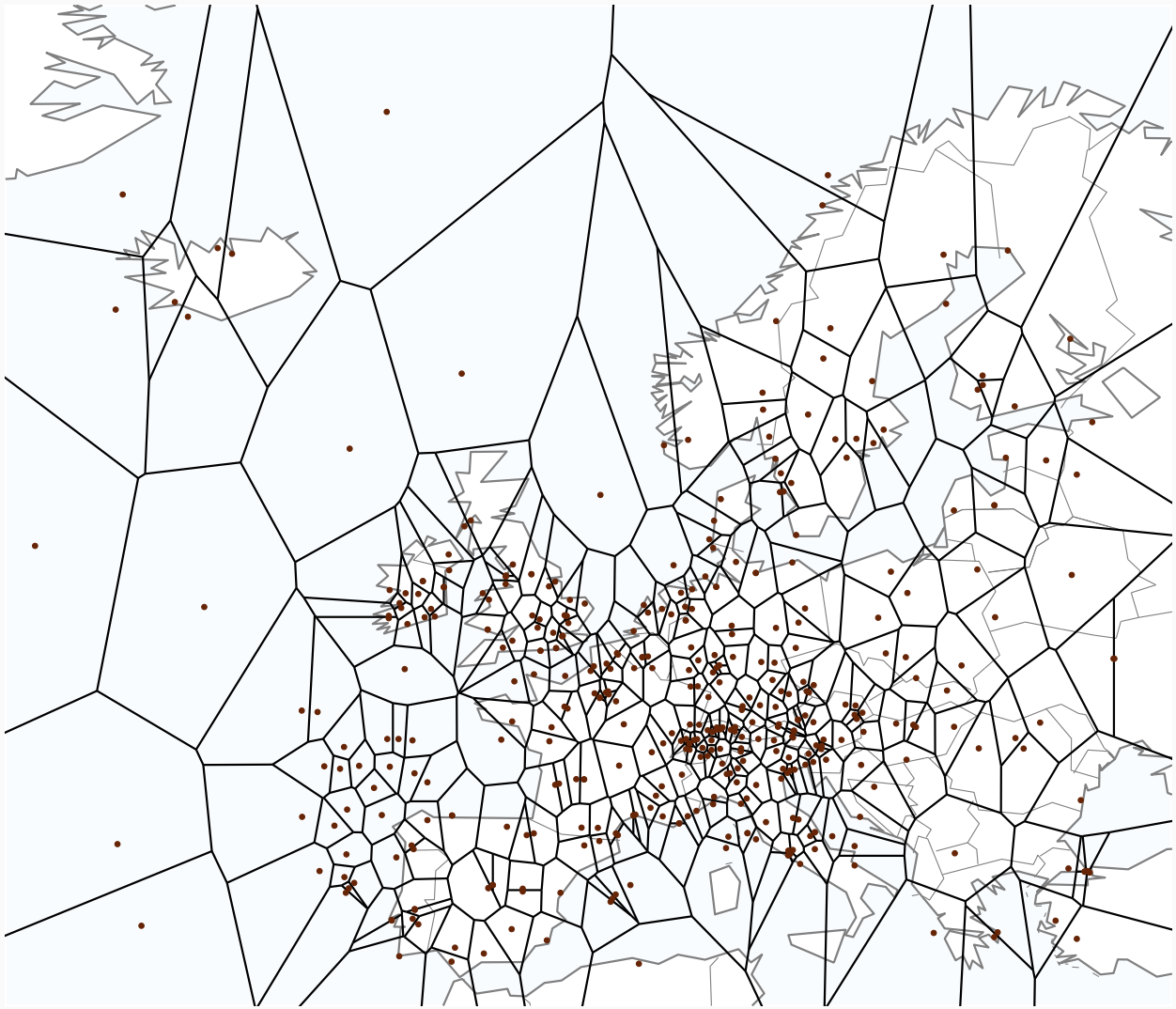
Aviation accidents



Aviation accidents: Voronoi diagram



Aviation accidents: Voronoi diagram



Contours and edges

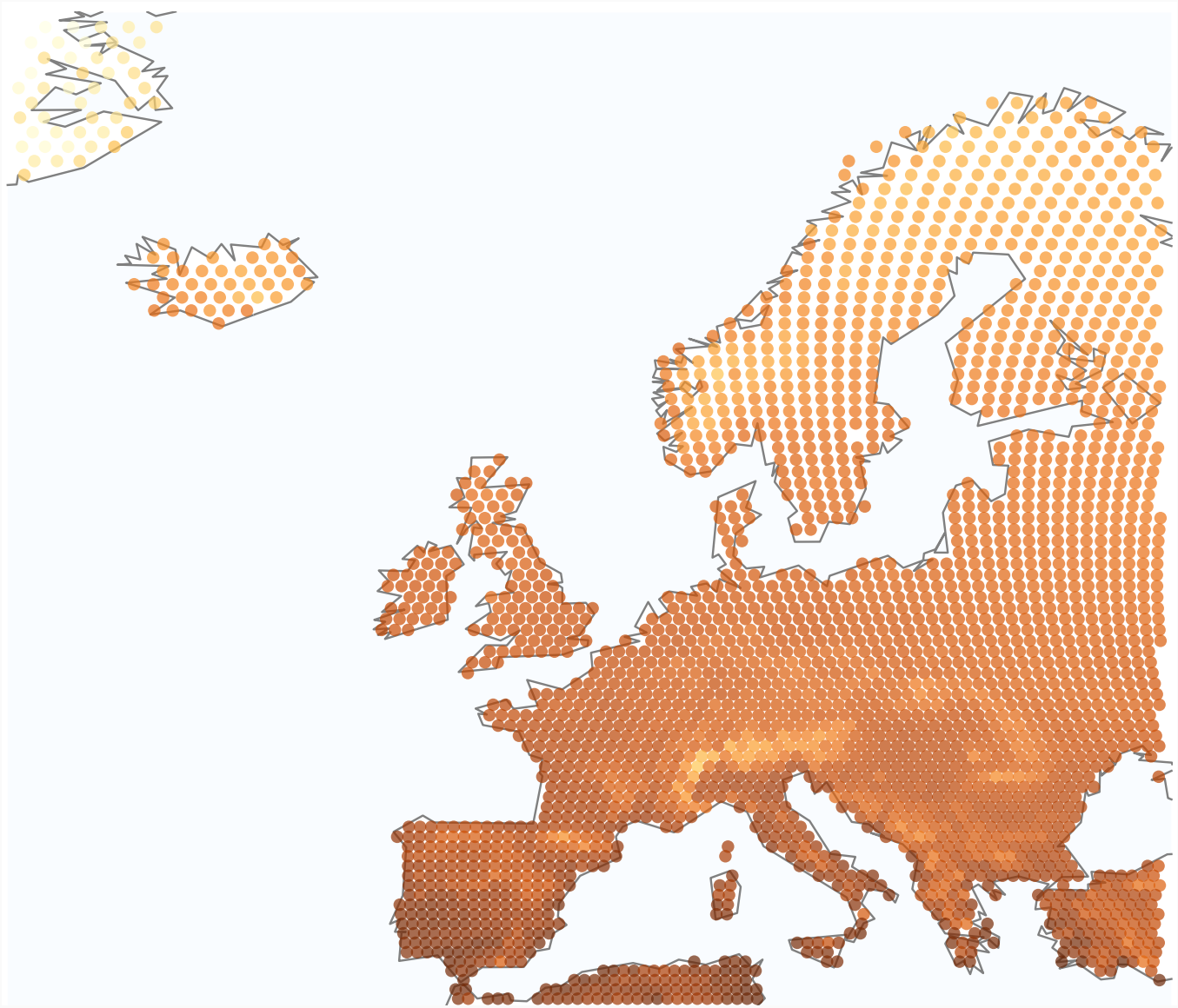
Compute value differences across neighboring points to identify areas at which value changes sharply

Edge detection methods aim at detecting points in an image at which value changes sharply

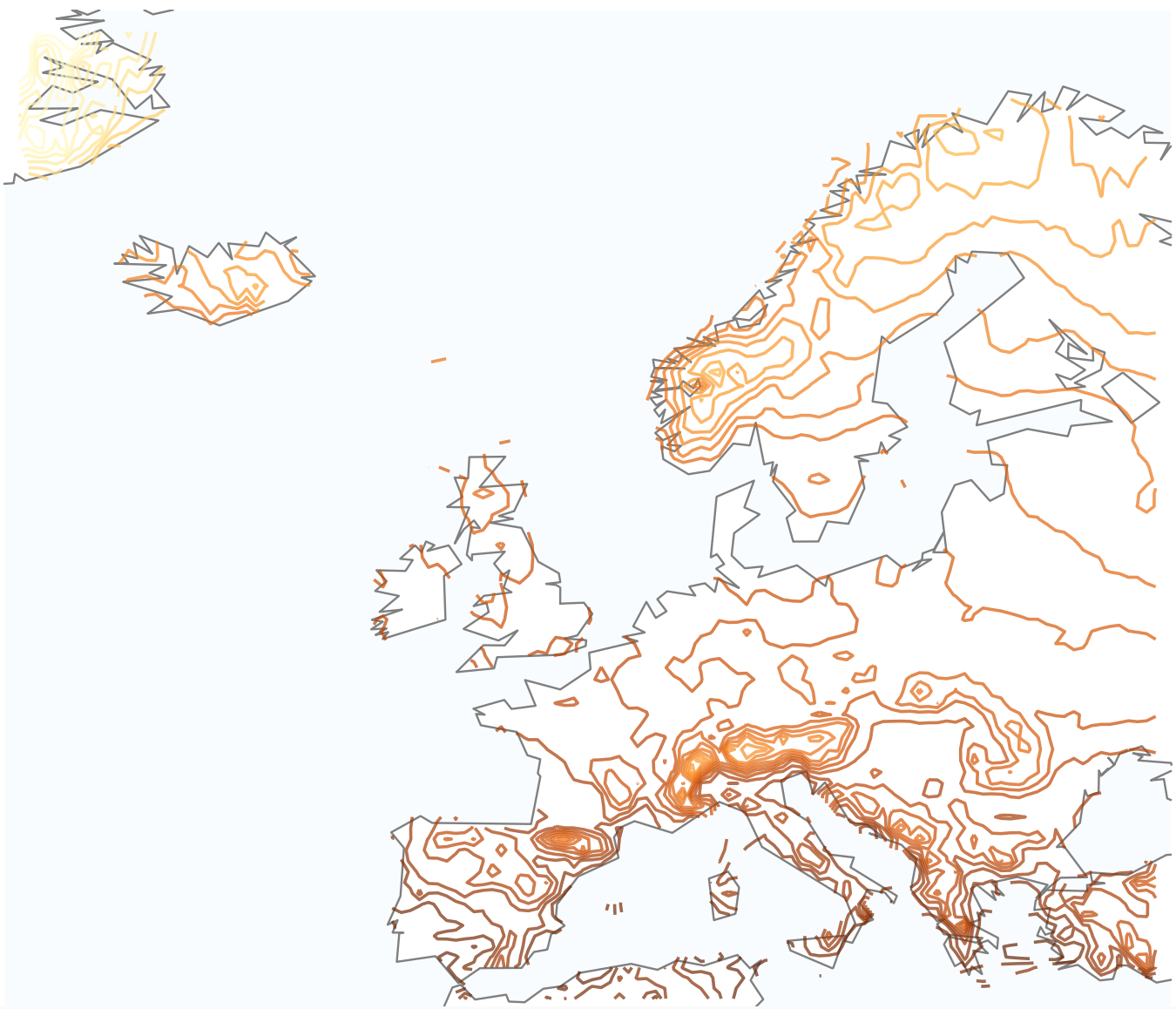
A *contour line* or *isoline* of a function of two variables is a curve along which the function has constant value

Consider a behavioral attribute as function of the coordinates
Contours are typically plotted for values spaced regularly across the domain of the attribute
Close contours indicate steep slopes, i.e. regions where the value of the attribute changes sharply

Yearly mean temperature



Yearly mean temperature



Shapes to time-series

Analysing shapes is challenging due to variations in size and orientation

The contour of a shape can be transformed into a time-series
Measure the distance from the centroid of the shape to its boundary, doing a clockwise sweep
E.g. taking 360 different regularly spaced angular samples produces a series of 360 numerical values

The time-series is referred to as the **centroid distance signature** of the shape

Shapes to time-series

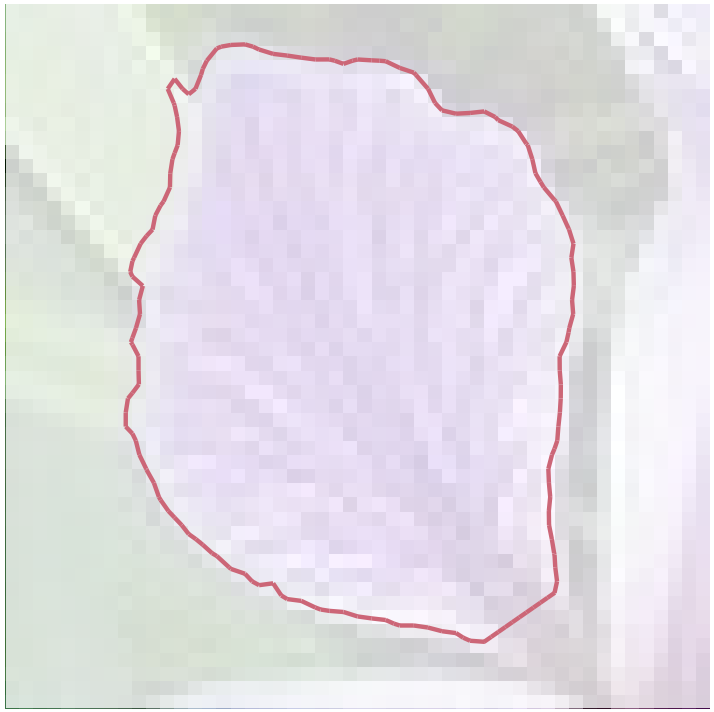
The contour of a shape can be transformed into a time-series
Measure the distance from the centroid of the shape to its boundary, doing a clockwise sweep
E.g. taking 360 different regularly spaced angular samples produces a series of 360 numerical values

The time-series is referred to as the **centroid distance signature** of the shape

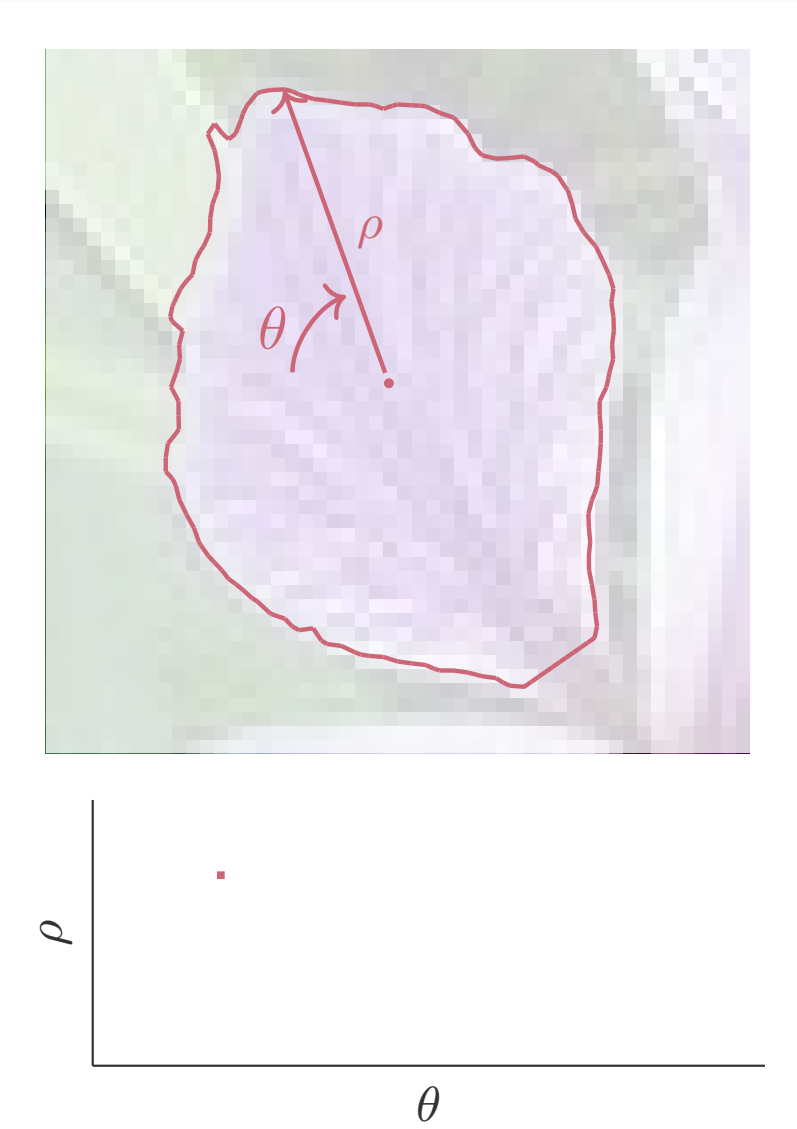
Rotations of the shape result in cyclic translation of the series
Mirror images of the shape result in a reversal of the series
Need to be taken into account in the analysis process

Shapes to time-series

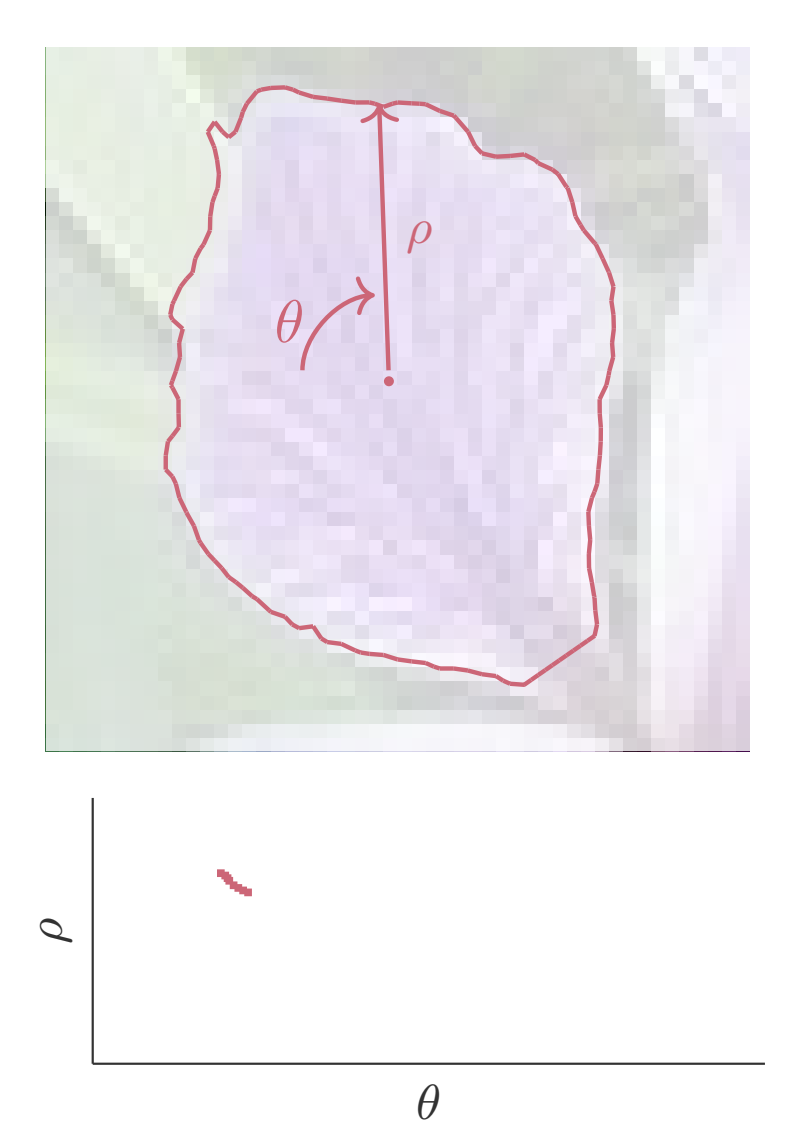
Shapes to time-series



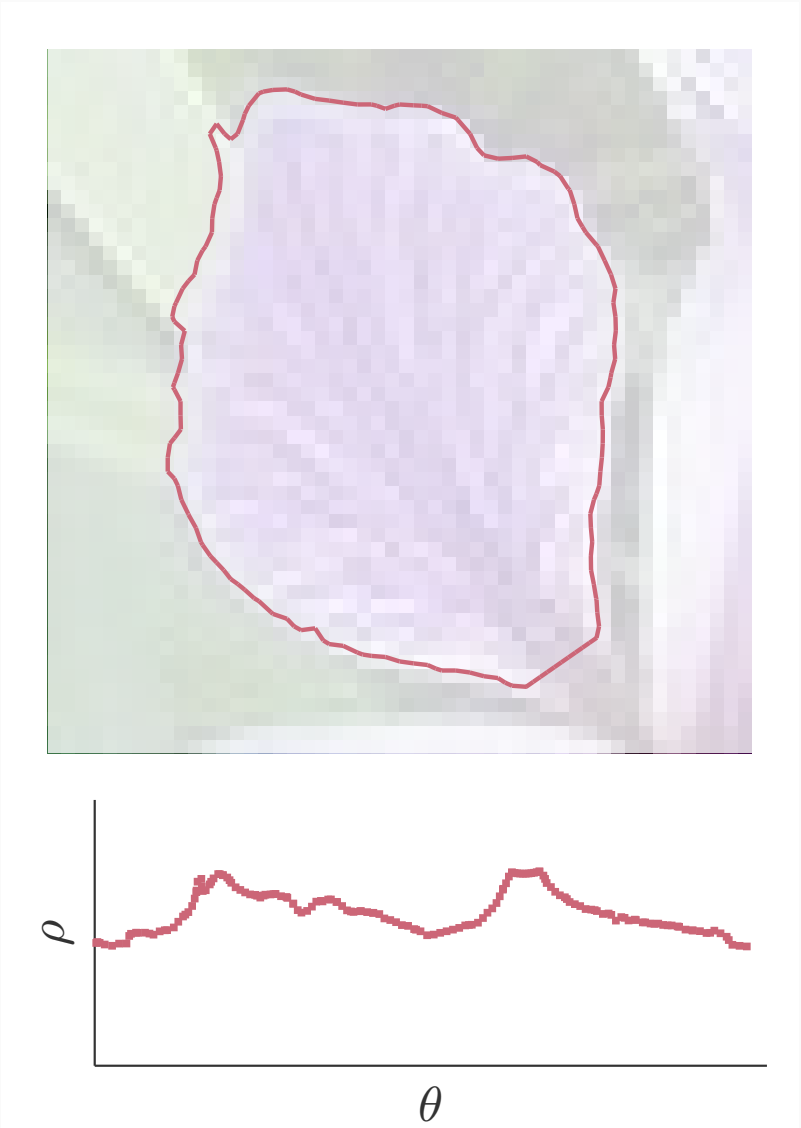
Shapes to time-series



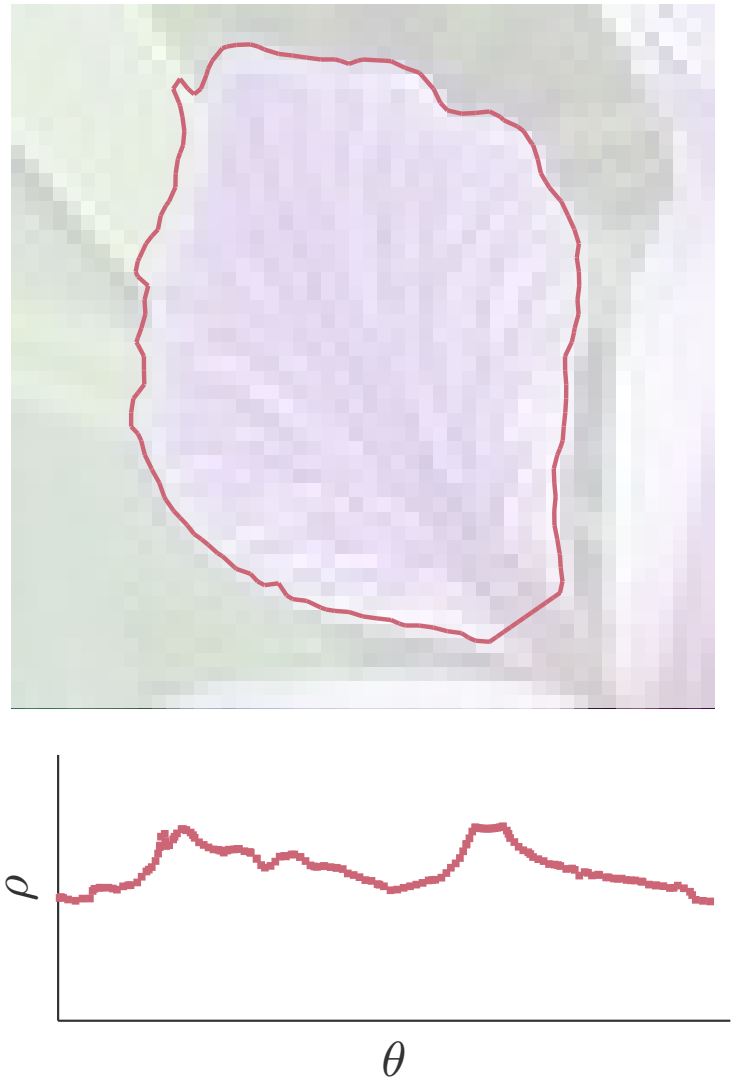
Shapes to time-series

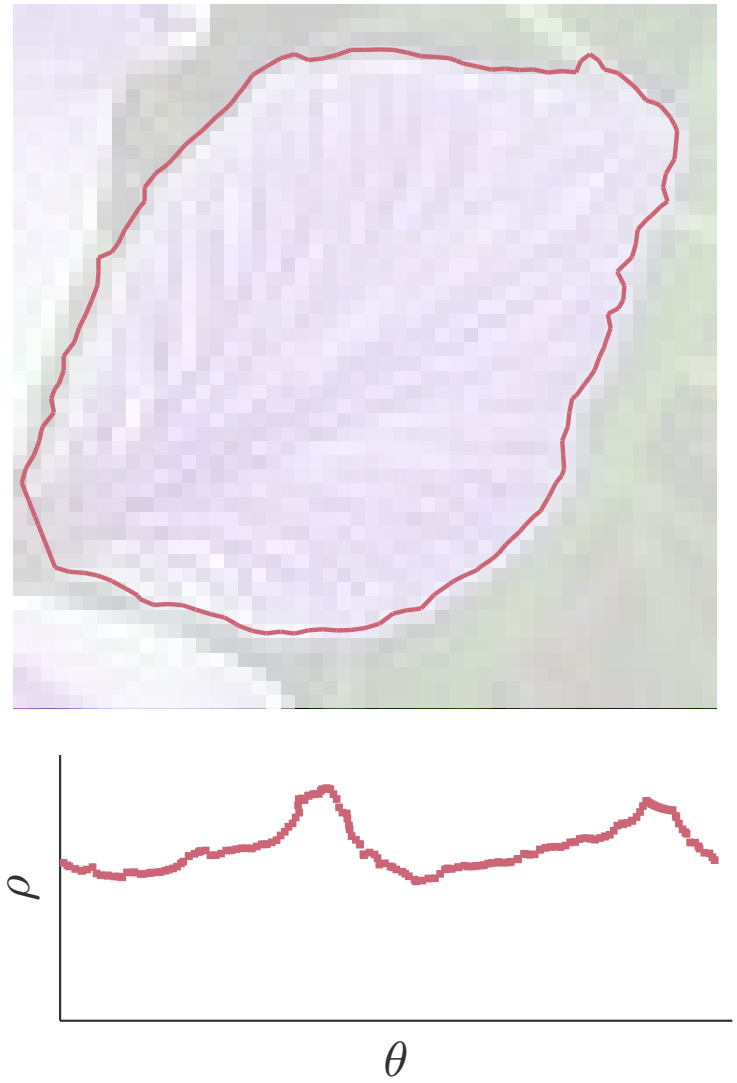


Shapes to time-series



Shapes to time-series





Discrete wavelet transform (DWT)

For numerical data arranged into a grid, spatially adjacent values are often very similar, storing all the values is wasteful, redundant

The discrete wavelet transform can be generalized to multiple contextual attributes
Differencing is applied across contiguous areas of the grid
Division is performed while alternating between the axes of the grid, i.e. the contextual attributes

Trajectory data

Object tracking

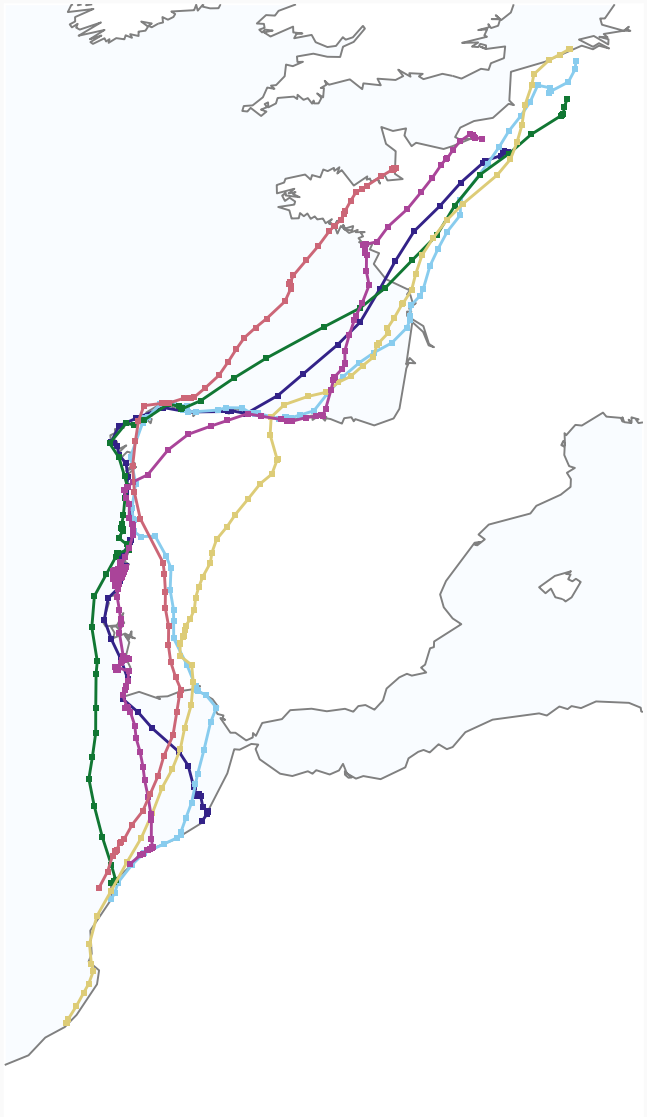
The position of a vehicule, robot, person, animal, etc. can be recorded over time through a variety of means, including the global positioning system (GPS), video, wireless triangulation, radio frequency identification (RFID)

A **trajectory** is a time-series of geo-locations
Time is the contextual attribute
Spatial coordinates constitute behavioral attributes

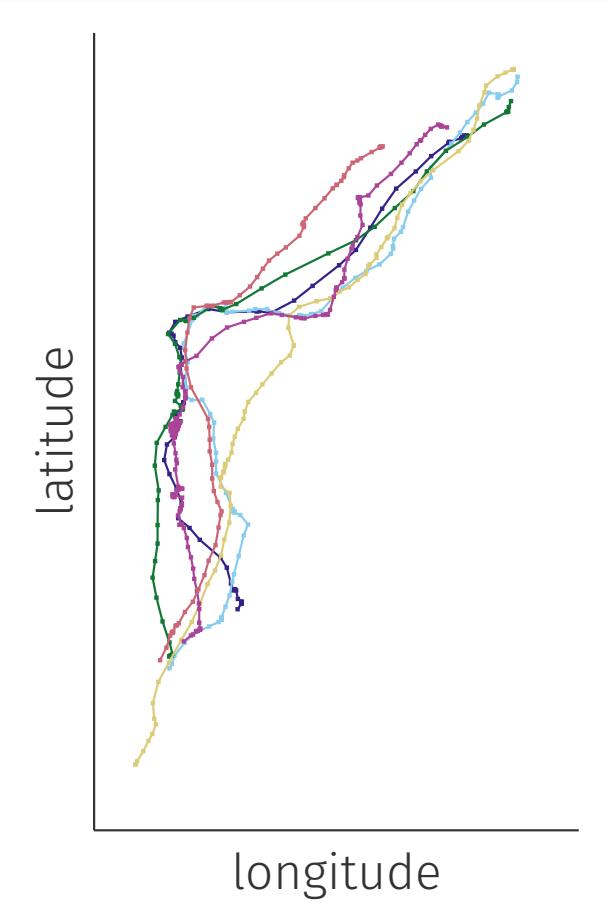
Transform a trajectory into multidimensional data

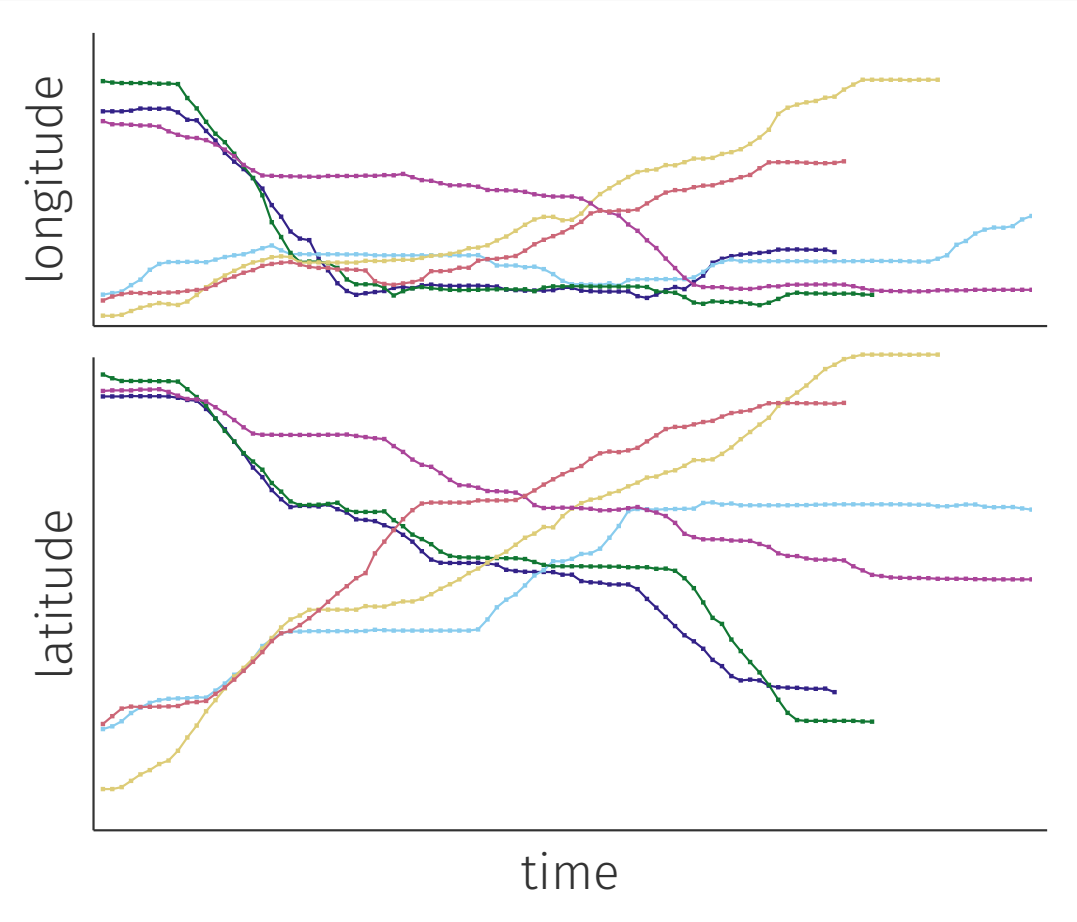
Compute the discrete wavelet transform coefficient for each spatial coordinate separately
Combine coefficients vectors across the different coordinates

Birds migration trajectories

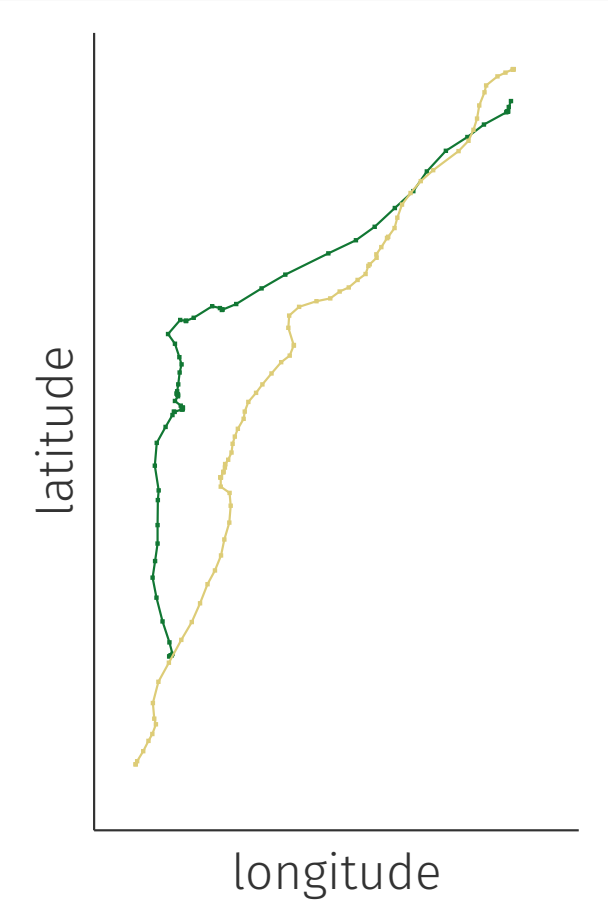


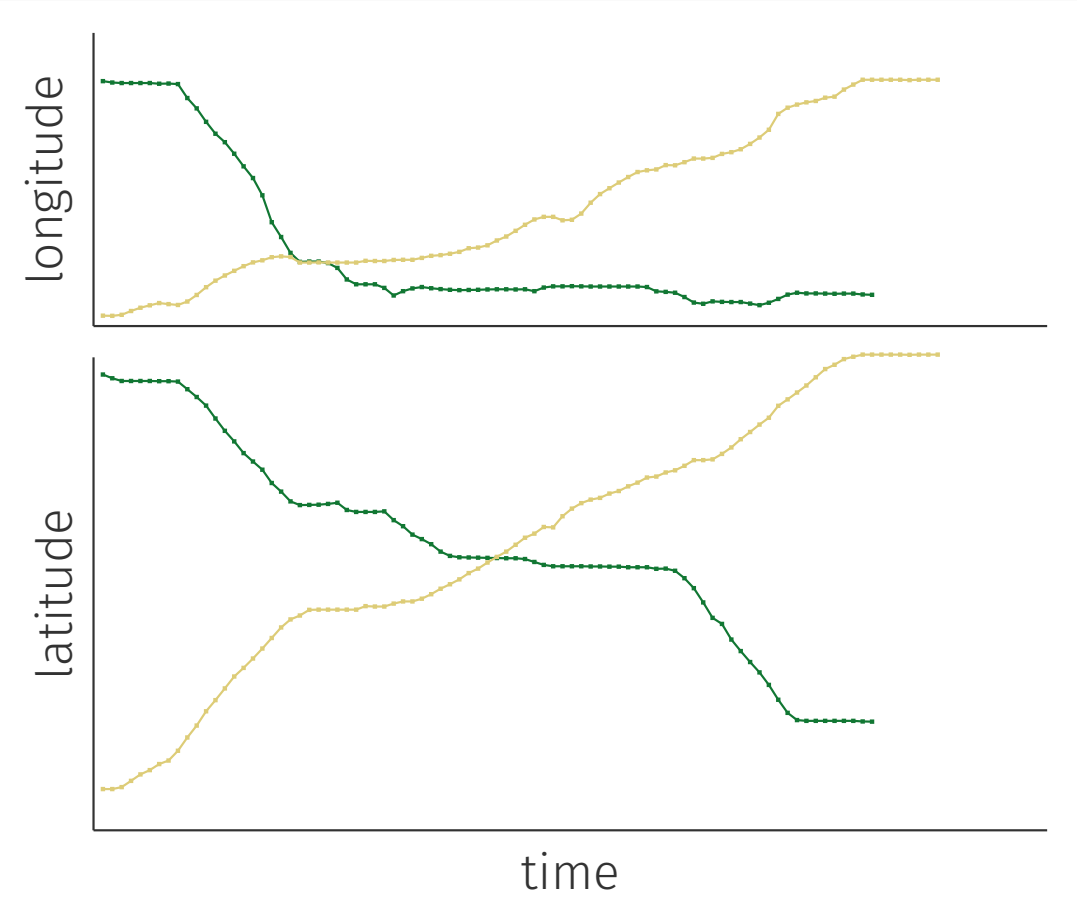
Birds migration trajectories as time-series





Birds migration trajectories as time-series





Trajectory data

Like other time-series, trajectories can be compared using the **dynamic time warping distance (DTW)**

$D_{DTW}(\mathcal{S}_X, \mathcal{S}_Y) = \text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(n_X, n_Y)$
where DTW is defined recursively

$$\text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(i,j) = \\ \mathsf{d}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)}) + \mathbf{min} \begin{cases} \text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(i, j-1) & \text{repeat } \boldsymbol{x}^{(i)} \\ \text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(i-1, j) & \text{repeat } \boldsymbol{y}^{(j)} \\ \text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(i-1, j-1) & \text{repeat neither} \end{cases}$$

with $\text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(0,0) = 0$,

$$\text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(i,0) = \infty, \forall i > 0 \text{ and } \text{DTW}_{\mathcal{S}_X,\mathcal{S}_Y}(0,j) = \infty, \forall j > 0$$

where $\mathsf{d}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(j)})$ is the distance between the position at time i in trajectory $\mathcal{S}_X$ and the position at time j in trajectory $\mathcal{S}_Y$

Frequent trajectory patterns

A key problem in analysing trajectories is to **identify frequent sequential paths**

1. Transform a trajectory into a univariate discrete sequence through grid-based discretization
2. Apply a sequential pattern mining algorithm (e.g. GSP) to the sequence(s)

Spatial tile transformation

Discretize each coordinate and assign a symbol to each interval

Each tile is identified by the combination of symbols along the different dimensions

Build the sequence associated to a trajectory by listing the identifiers of the tiles it traverses

Frequent trajectory patterns

A key problem in analysing trajectories is to **identify frequent sequential paths**

1. Transform a trajectory into a univariate discrete sequence through grid-based discretization
2. Apply a sequential pattern mining algorithm (e.g. GSP) to the sequence(s)

Spatio-temporal tile transformation

Divide the time range into intervals and assign them identifiers
For a given trajectory, list for each time interval the identifiers of the tiles in which at least a chosen amount of the interval was spent, tagged with the corresponding interval identifier

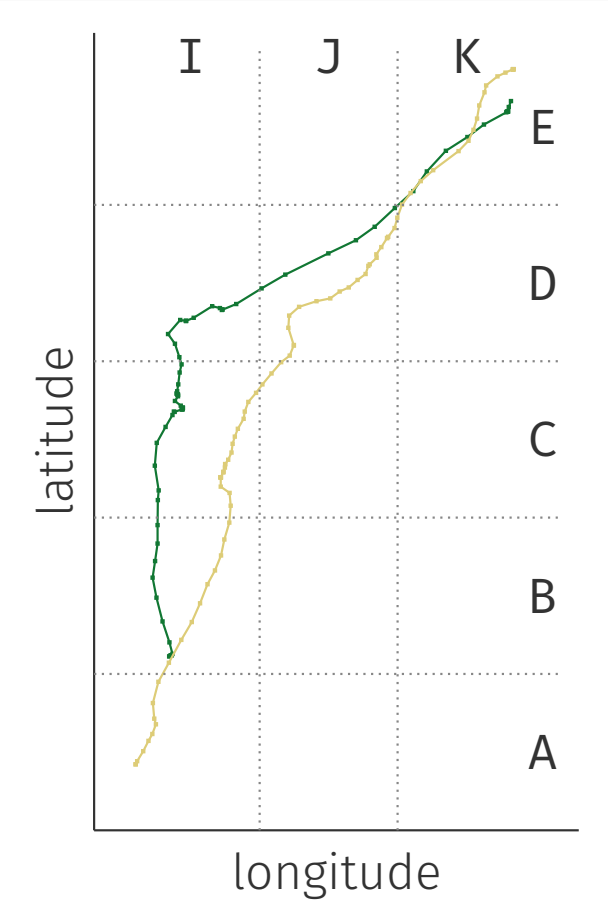
Frequent trajectory patterns

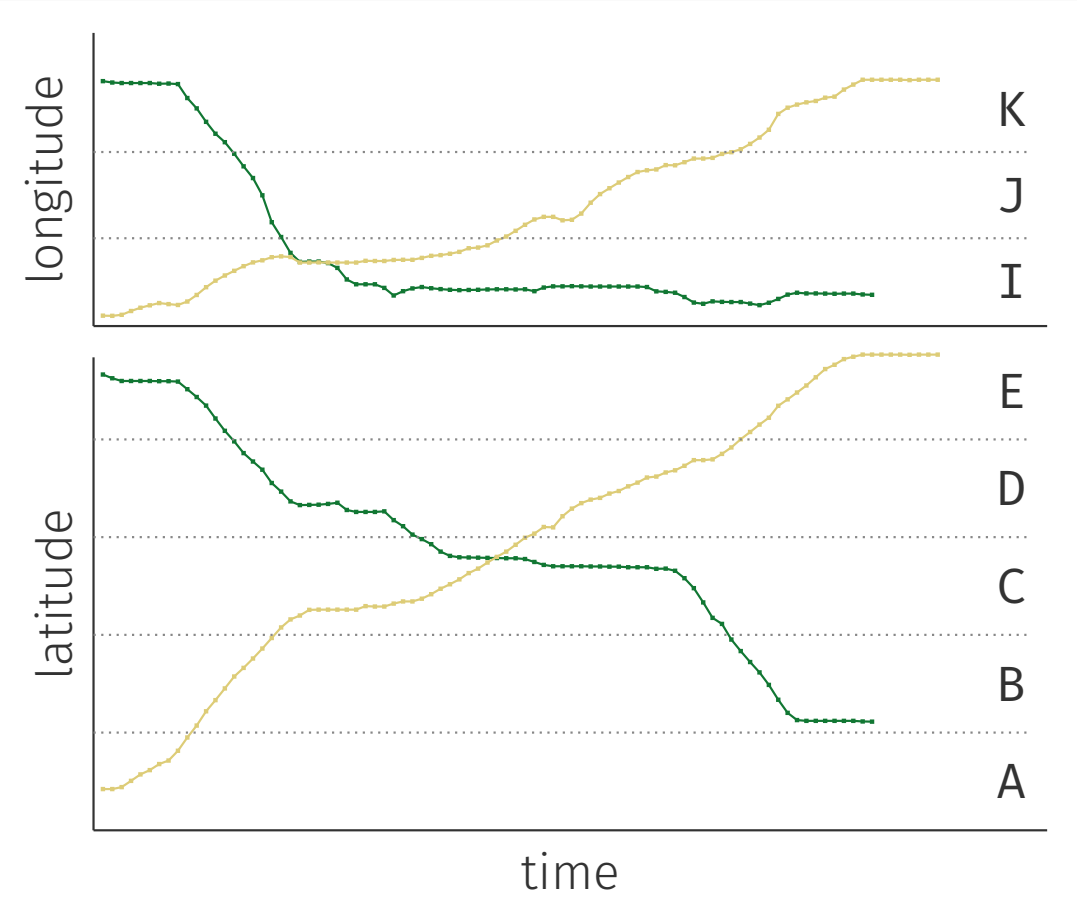
Spatio-temporal tile transformation

Divide the time range into intervals and assign them identifiers
For a given trajectory, list for each time interval the identifiers of the tiles in which at least a chosen amount of the interval was spent, tagged with the corresponding interval identifier

Allows to identify simultaneous movements across different trajectories and affords increased flexibility
The granularity of the discretization might be difficult to adjust and can impact the results

Birds migration trajectories: spatial tile transformation





Birds migration trajectories: spatio-temporal tile transformation