1 Distance Calculation

In this section we will discuss the various parameters and methodologies of calculating each type of distance separately and then explain how these distances are combined to give a composite distance value between two events.

1.1 Spatio-temporal Distance

To combine the spatial and temporal distances into a spatio-temporal distance, we use a weighted sum of these two distances. In order to do so, we introduce the notion of the primary event and the secondary event: a primary event occurred before a secondary event. In case of the two events occurring on the same date, the event that occurred inside a geographic region with higher population is tagged as the primary, creating an ordered temporal relationship between all pairs of events. We also consider population sizes in the design of the weights. We see that time has less value in loosely populated area than it has in a densely-populated one. This is based on the hypothesis that an unrest event would take longer to evolve or spread in a sparsely populated area than it would in an area with high population. For example, in a sparsely populated village, it is normal for information to take a longer time to travel, whereas in cities where the population is dense, it would take a much shorter time for information to travel. We therefore argue that a larger temporal difference between events in a village is equivalent to a much shorter temporal difference in a city. At the same time, if we look at the spatial distance, an unrest event will most likely spread much faster in a very densely populated area than it does in a loosely populated area. For example, a 5-km radius in a city is a smaller distance compared to a 5-km radius in a village, and so a 5-km distance in a city might be equivalent to a 2-km distance in a village. Given this, we have the following weight assignment scheme for the temporal weight (w_t) and spatial weight (w_s):

$$w_t = \frac{population(event\_primary)}{population(event\_primary) + population(event\_second)} \tag{1}$$

$$w_s = 1 - w_t \tag{2}$$

The calculation of spatial distance is relatively straightforward, but since we are working with social events that have a time range instead of a single date value, we cannot simply check the difference in dates to calculate the temporal distances. We therefore look at two different temporal metrics to calculate the final temporal distance.

Temporal Gap

The gap metric can be defined as the time period between any two events compared to the total temporal span created by the start of an event to the end of a later event. Overlapping events have negative gap metrics. The gap metric allows us take the event duration into account when looking at the temporal

gap of between events. If e_1 and e_2 are two events with $(t_{e_1,start}, t_{e_1,end})$ as the start and end dates of event e_1 and $(t_{e_2,start}, t_{e_2,end})$ as the start and end dates of event e_2, the temporal span of a pair of events is defined as the time period from the minimum start-date of the pair of events, to the maximum end-date. The temporal span is calculated as:

$$span = days\_between[\max(t_{e_1,end}, t_{e_2,end}), \min(t_{e_1,start}, t_{e_2,start})] \tag{3}$$

and the gap between events is calculated as:

$$gap = days\_between[\max(t_{e_1,start}, t_{e_2,start}), \min(t_{e_1,end}, t_{e_2,end})] \tag{4}$$

The gap metric is then given by:

$$\Delta t_{gap}(e_1, e_2) = \frac{gap}{span} \tag{5}$$

Temporal Coverage

While the gap metric gives us a good estimate of the distance between two events, it is solely based on the gap between events and the overall temporal span, it does not take the coverage or range of individual event into account. We therefore introduce a coverage metric, which is calculated as the relative difference between the two time periods using a Euclidian distance metric by treating the start and end dates of any events as the x and y coordinates. The value is then divided by the Manhattan distance, this is done simply to normalize the Euclidean metric. If e_1 and e_2 are two events with $t_{e_1,start}$ and $t_{e_2,start}$ as their start dates, and $t_{e_1,end}$ and $t_{e_2,end}$ as their end dates respectively, the separation metric is given by:

$$\Delta t_{Euclidean}(e_1, e_2) = \sqrt{(t_{e_1,end} - t_{e_2,end})^2 + (t_{e_1,start} - t_{e_2,start})^2} \tag{6}$$

$$\Delta t_{Manhattan}(e_1, e_2) = \max(|t_{e_1,end} - t_{e_2,end}| + |t_{e_1,start} - t_{e_2,start}|, 1) \tag{7}$$

$$\Delta t_{coverage}(e_1, e_2) = \begin{cases} \dfrac{\Delta t_{Euclidean}}{\Delta t_{Manhattan}} & \text{if } \Delta t_{Manhattan} > 0, \\ 1 & \text{if } t_{Manhattan} = 0 \end{cases} \tag{8}$$

The final temporal distance between any pair events e_1 and e_2 is then calculated as:

$$d_{temporal}(e_1, e_2) = w_{gap}.\Delta t_{gap}(e_1, e_2) + w_{coverage}.\Delta t_{coverage}(e_1, e_2) \tag{9}$$

where, w_{gap} and $w_{coverage}$ are weights assigned to the temporal overlap metric and the temporal separation metric, respectively, such that: $w_{overlap} + w_{coverage} = 1$.

Finally, we combine the spatial and temporal distance together to give a spatio-temporal distance. Since the spatial and temporal distances are in different numeric ranges, we normalize the spatial distance. The normalization of

spatial distance is done by dividing the spatial distance with a threshold value ($d_{spatial\_threshold}$) beyond which the spatial distance is assumed to have the same meaning (too far apart). The normalized spatial distance is then calculated as:

$$d'_{spatial} = \min\left(\frac{d_{spatial}}{d_{spatial\_threshold}}, 1.0\right) \tag{10}$$

The spatio-temporal distance between events e_1 and e_2 is given by:

$$d_{spatio-temporal}(e_1, e_2) = w_s.d'_{spatial}(e_1, e_2) + w_t.d_{temporal}(e_1, e_2) \tag{11}$$

Where, $d'_{spatial}(e_1, e_2)$ is the normalized value of the spatial distance, calculated by w_s and w_t, which are the spatial and temporal weights, respectively.

1.2 Socio-Economic Distance

To calculate the socio-economic distance between two events, we first take the difference in the values of the same type of socio-economic variable assigned to both events. For any unrest event e, we compute a corresponding vector of socio-economic variables: $\langle v_{1,e}, , v_{n,e}\rangle$, where $v_{k,e}$ represents the k^{th} socio-economic variable of the event e that has been normalized so that all the variables are comparable. Let wt_k be the weight applied to k^{th} variable based on its relative importance compared to other socio-economic variables. The difference between the k^{th} socio-economic variable of events e_1 and e_2 is given by:

$$\Delta_{socio-economic,k}(e_1, e_2) = |v_{k,e_1} - v_{k,e_2}| \tag{12}$$

The socio-economic distance between two events e_1 and e_2 is then calculated as:

$$d_{socio-economic}(e_1, e_2) = \sum_{k=1}^{n} wt_k.\Delta_{socio-economic,k}(e_1, e_2) \tag{13}$$

Since the weight is relative, $wt_1 + wt_2 + ... + wt_n = 1$.

1.3 Infrastructural Distance

As alluded to earlier, incorporating geospatial objects such as infrastructural elements is a challenge. Now, let us first suppose we have data for n different types of infrastructure elements (schools, hospitals, police-stations, etc.). We can define a vector of normalized distances of the nearest infrastructure element from event e by $i_{1,e}, ..., i_{n,e}$. The difference in $infrastructure\_proximity$ between two events e_1 and e_2 for the k^{th} infrastructure type is given by:

$$\Delta_{infrastructure\_proximity,k}(e_1, e_2) = |i_{k,e_1} - i_{k,e_2}| \tag{14}$$

For two events e_1 and e_2, the $infrastructure\_proximity$ distance can then be calculated as:

$$d_{infrastructure\_proximity}(e_1, e_2) = \sum_{k=1}^{n} wi_k.\Delta_{infrastructure\_proximity,k}(e_1, e_2) \tag{15}$$

We then collect the number of infrastructure elements in a pre-determined radius of an events location, separately for each infrastructure type. For any event e, the normalized count of infrastructure elements of n different types, within a pre-determined geospatial radius r can be denoted by $c_{1,e}, ..., c_{n,e}$. The difference in $infrastructure\_density$ between two events e_1 and e_2 for the k^{th} infrastructure type is given by:

$$\Delta_{infrastructure\_density,k}(e_1, e_2) = |c_{k,e_1} - c_{k,e_2}| \tag{16}$$

For two events e_1 and e_2, the $infrastructure\_density$ distances can then be calculated as:

$$d_{infrastructure\_density}(e_1, e_2) = \sum_{k=1}^{n} wi_k.\Delta_{infrastructure\_density,k}(e_1, e_2) \tag{17}$$

where, wi_k is the relative weight assigned to each individual infrastructure-type based on its importance. Since wi_k is relative, $wi_1 + ... + wi_n = 1$.

Similarly, we can define a vector of normalized distances of the nearest infrastructure element from event e, for n different types of infrastructures by $i_{1,e}, ..., i_{n,e}$. The difference in $infrastructure\_proximity$ between two events e_1 and e_2 for the k^{th} infrastructure type is given by:

$$\Delta_{infrastructure\_proximity,k}(e_1, e_2) = |i_{k,e_1} - i_{k,e_2}| \tag{18}$$

For two events e_1 and e_2, the $infrastructure\_proximity$ distance can then be calculated as:

$$d_{infrastructure\_proximity}(e_1, e_2) = wi_k.\Delta_{infrastructure\_proximity,k}(e_1, e_2) \tag{19}$$

We will then combine these distances to give the comprehensive infrastructure distance between two events by assigning weights to the distances calculated above.

$$d_{infrastructure}(e_1, e_2) = w_{density}.d_{infrastructure\_density}(e_1, e_2) + w_{proximity}.d_{infrastructure\_proximity}(e_1, e_2) \tag{20}$$

where, $w_{density}$ and $w_{proximity}$ are weights based on the relative importance of the types of distances, such that $w_{density} + w_{proximity} = 1$.

1.4 Integrated, Multi-factorial Distance

The final distance between events e_1 and e_2 is given by the weighted sum of spatio-temporal, socio-economic and infrastructure distances.

$$d(e_1, e_2) = d_{socio-economic}(e_1, e_2).w_{se} + d_{spatio-temporal}(e_1, e_2).w_{st} + d_{infrastructure}(e_1, e_2).w_{in} \tag{21}$$

where w_{se} is the socio-economic weight, w_{st} the spatio-temporal weight, and w_{in} the infrastructure weight.