Heterogeneous Information Fusion - Case Study using Visual Analytics Science and Technology (VAST) 2018 Challenge Dataset

Paweł Kowalski

June 15, 2018

Abstract

The IEEE Visual Analytics Science and Technology (VAST) Challenge is an annual contest with the goal of advancing the field of visual analytics. The VAST Challenge problems provide researchers with realistic tasks and data sets for evaluating their software, as well as an opportunity to advance the field by solving more complex problems. The dataset associated with a subset of the 2018 Challenge (Mini-Challenge 1) is used to demonstrate and showcase heterogeneous multi-source fusion, by using deep learning, machine learning, pattern-of-life analysis and traditional statistical methods to model knowledge and uncertainty in the framework of belief functions (also known as theory of evidence or Dempster-Shafer theory). The variables of interest are defined and the interrelationships between themselves and the contributing sources of evidence (belief assignments) are modelled using an informal graphical framework. The information sources are combined using appropriate combination rules and the provenance of belief is tracked using the graph model in conjunction with a contribution measure.

1 Challenge description

In this case study we attempt to answer the questions stated in VAST mini-challenge 1 from 2018. The following is a shortened backstory from VAST website. All names and locations are fictional.

he IEEE Visual Analytics Science and Technology (VAST) Challenge is an annual contest with the goal of advancing the field of visual analytics. The VAST Challenge problems provide researchers with realistic tasks and data sets for evaluating their software, as well as an opportunity to advance the field by solving more complex problems. In 2017, the VAST Challenge results suggested that the Kasios Furniture manufacturing company may have been a primary contributor to the apparent reduction of the number of nesting pairs of the Rose-Crested Blue Pipit (RCBP). Kasios supposedly used the banned substance Methylosmolene in their manufacturing process. They dumped process waste in the northeast region of the Preserve and Methylosmolene was detected in their smokestack emissions. Kasios now claims that the analysis was flawed and biased. To back up this claim, they have provided a set of Pipit bird calls, recently recorded across the Preserve, with locations of where they were recorded. Another collection of bird calls from the preserve that has been vetted by various ornithology groups as having accurate identifications is provided along with corresponding geospatial information and a timestamp.

The problem to be answered is as follows:

1. Using the bird call collection and the map of the Wildlife Preserve, characterize the patterns of bird species over the time of collection (as per official problem statement). Here what we want to focus on in particular are two major questions - is the population of the RCBP waning? And is there a specific pattern in the behaviour of the RCBP which can be attributed to the location where Kasios is dumping the process waste?

2. Does the set of bird calls provided by Kasios support their claim that Pipits are being found across the Preserve? Here we are not concerned with the statistical impact of sample size of the test set, but rather try to answer the following questions:
    - Are the locations where the pipits were recorded consistent with the patterns of life determined in (1)?
    - Are the actual recordings provided by Kasios consistent with the RCBP recordings from the vetted dataset?
3. Formulate a hypothesis concerning the state of the RCBP and whether Kasios is responsible to it - whilst providing a clear track of evidence used to support the conclusion.

1.1 Dataset overview

The dataset consists of the following

- A map of the Wildlife Preserve and the location of the Kasios process waste dump site
- Library of 2082 bird recordings im mp3 format with varying lengths each with the corresponding metadata: label as call or song, label by bird type (including 187 RCBP labels), quality rating, X and Y coordinates of the location where the recording was taken and a timestamp
- 15 recordings labeled as RCBP provided by Kasios with X and Y location coordinates

1.2 Situation model

In order to model the situation we define the following variables:

- L - Recording locations provided by Kasios are consistent with RCBP patterns of life
- R - The recorded sounds (provided by Kasios) are typical for RCBP
- K - Files provided by Kasios are legitimate
- P - the population of RCBP is dwindling
- D - the process waste dump site has an impact on RCBP behaviour
- F - Kasios is to be blamed for RCBP problems

We can define relations between these variables in natural language and then formally represent them using propositional logic

1. The files provided by Kasios are legitimate if both the geographical locations are consistent with RCBP patterns of life and the recordings contain RCBP sounds $L \wedge R \rightarrow K$
2. If the locations are inconsistent or the recordings do not contain RCBP sounds it implies that the Kasios files have been falsified $\bar{L} \vee \bar{R} \rightarrow \bar{K}$
3. If Kasios files are legitimate the RCBP population is not dwindling $K \rightarrow \bar{P}$
4. If the RCBP population is dwindling and RCBP patterns of life are affected by the dumping site Kasios is to be blamed for that $P \wedge D \rightarrow F$

Finally we can define one additional relation (is it fuzzy logic?) - note different arguments could be made using game theory

5. If Kasios files are deemed fraudulent it partially implies that Kasios is guilty (i.e. it provides some evidence δ towards that hypothesis). Here we model $\delta = 0.4$ and thus $\bar{K} \overset{0.4}{\dashrightarrow} F$

These relations are represented in graphical form in Figure 1.

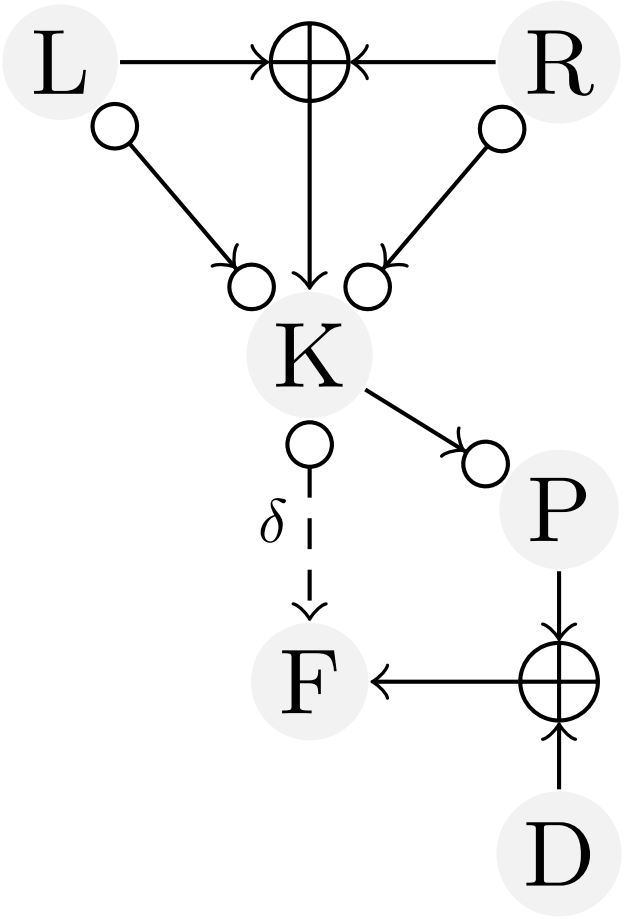

Figure 1: Graphical situation model

2 Data analysis

This section discusses the methods used to analyse data and generate basic belief assignments which later feed into the situation model as shown in Figure 1. An important disclaimer to be made at this point is that the selection of methods used both to analyse the data and to model uncertainty associated with the results is not necessarily optimal - it is a complex problem and multiple solutions exists; throughout this example it is often the case that a simpler model is selected in favour of a more accurate one in order to avoid increasing complexity (this is in particular true for generation of basic belief assignments) .

2.1 Analysis of recordings

The purpose of analysis of recordings themselves is to determine whether the recordings are in fact recordings of RCBPs or rather of some other birds or synthetic. In order to do that we represent the sound data in an image form and then analyse it using a variety Convolutional Neural Network (CNN) techniques.

First of all we convert the mp3 files to 2-dimensional spectrograms using the *specgram* function from *matplotlib.pyplot*. A typical output resembles the one in Figure 2

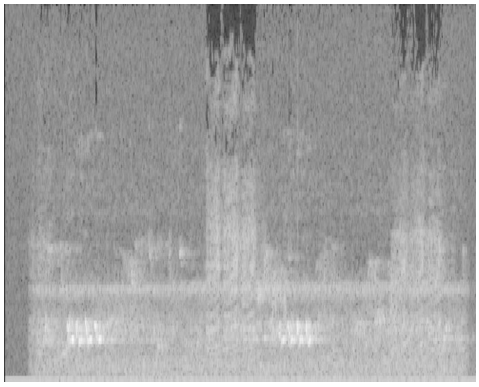

Figure 2: A typical greyscale spectrogram of a RCBP recording

Using this we attempt to answer two questions

1. Are the recorded birds pipits (as opposed to another bird type or a synthetic recording)?
2. Are the pipit recordings typical for pipit recordings?

In order to answer the former we train a CNN classifier on the vetted dataset and use the identification rate of the Kasios bird set as a metric; in order to answer the latter we use a convolutional autoencoder and compate the reconstruction error of the RCBP recordings from the vetted dataset to that from the Kasios test dataset.

The classification case is implemented with a simple architecture with 3 convolutional layers, each of them followed by a pooling layer. This is then flattened; followed by a dense layer, a dropout layer for training and a logit layer. The Tensorflow implementations of both that and the autoencoder are available on Github (ref).

The network is independently trained twice, both times using 80% of the samples from the vetted dataset for training and the remainder for verification. The performance metrics from both runs are displayed in Table 1. An interesting point to note is that in both cases the classification accuracy of RCBP is above average

Run	Mean accuracy	Mean per class accuracy	RCBP accuracy	Number of training steps
1	0.4375	0.3567	0.6	5000
2	0.4567	0.4039	0.77	8000

Table 1: Performance of CNN classifiers

Once training is complete predictions are made on the test dataset provided by Kasios. The results show very low rate of RCBP identification: 1/15 in run 1 and 2/15 in run 2.

The second method of analysis we employ aims to assess how similar to other pipit recordings are the ones provided by Kasios. In order to do that a convolutional autoencoder (an encoder followed

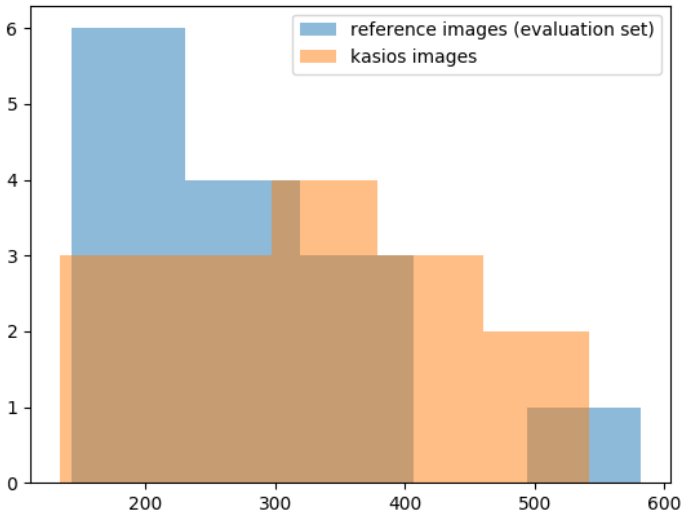

Figure 3: A histogram showing typical reconstruction error for a single iteration of the train-and-evaluate loop

by a decoder) is implemented. Such an architecture attempts to reconstruct the original image and as such learns the typical patterns present in images it is reconstructing; therefore the reconstruction error can be used as a measure of "distance from mean" for specific input. The encoder is built from two convolutional layers, each followed by a pooling layer; the decoder is built from two upsample layers each followed by a convolutional layer, the output of the final one is fed to a logits layer.

The important distinction here is that under this approach the neural network is trained only on RCBP recordings and as such the training dataset is much smaller (187 entries only). In order to draw conclusions regarding statistical dissimilarity between the Kasios and reference datasets we proceed as follows:

1. A training sample of 93 elements is randomly selected and the network is trained (2000 steps, batch of 4)
2. Reconstruction error values are computed for the remaining elements in the RCBP reference dataset as well as the 15 files provided by Kasios
3. This is performed 10 times and the results are aggregated

While this will produce suboptimal reconstruction error values for the reference (test) set, it ensures that overfitting is avoided. A typical set of results for a single iteration of the above algorithm is shown in Figure 3

Aggregated results are shown in the following two figures with the histogram of accumulated figures in Figure 4 and the fitted Epachenikov kernel CDF in Figure 5. Without drawing further conclusions at this stage it can easily be seen that the reconstruction error on the test recordings provided by Kasios is, on average, larger than reference, suggesting that there is some inherent difference between these two sets.

2.2 Patterns of life analysis

Through patterns of life analysis we attempt to answer the following questions:

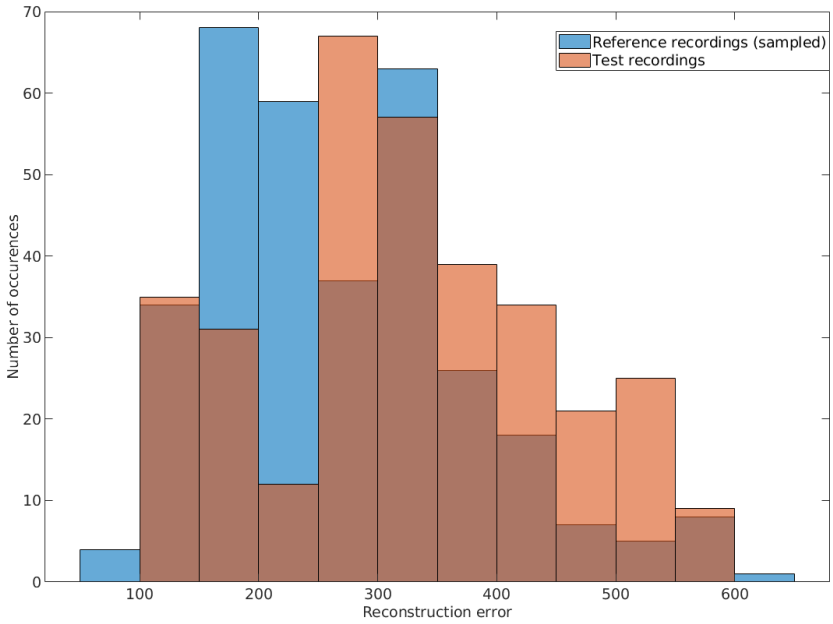

Figure 4: Histogram showing aggregated reconstruction error for multiple iterations of the train-and-evaluate loop

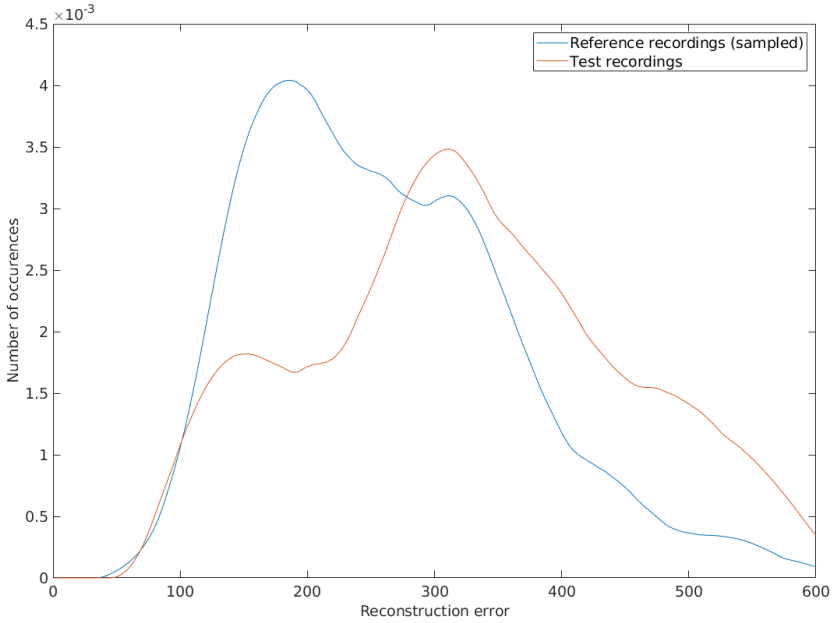

Figure 5: Epachenikov kernel CDF fitted to the reconstruction error for multiple iterations of the train-and-evaluate loop

- To what extent the geospatial coordinates provided by Kasios regarding the recording locations are consistent with patterns of life of RCBP?
- Has there been a significant decrease in RCBP population in recent years?

- Have RCBP patterns of life been affected by Kasios dumping process waste?

In order to answer the first question stated we use classical machine learning methods to identify anomalies in the joint distribution of the test and reference RCBP recording locations. This geospatial distribution is shown in Figure 6. In order to avoid the issues caused by bird migration and long-term patterns of life we only compare the test data provided by Kasios to the RCBP recordings obtained from 2015 onwards.

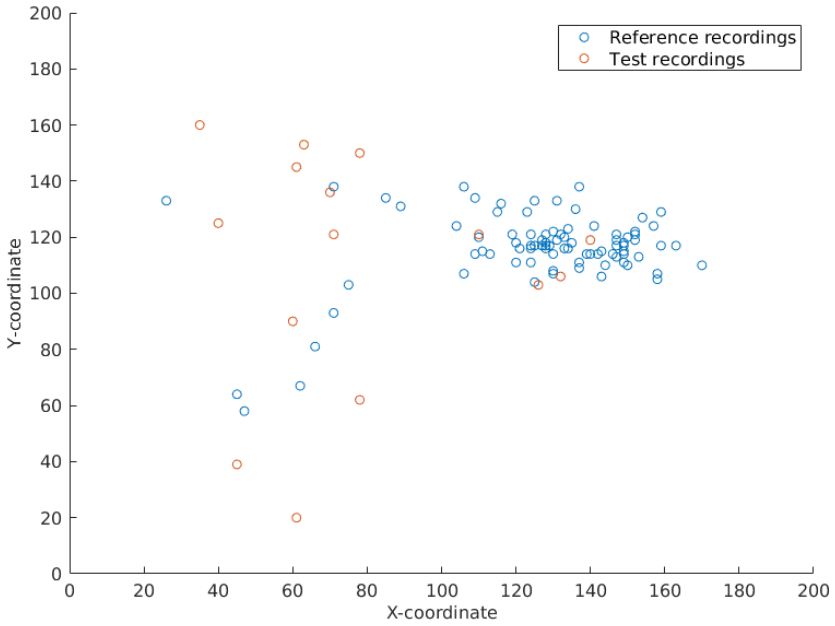

Figure 6: Locations provided for test and reference recordings

A simple method of identifying anomalies in an unsupervised manner is the K-means clustering algorithm. In a standard anomaly detection situation, anomalous elements would end up in low population clusters. However, since our application is somewhat different as we are comparing two sets of samples, we can measure dissimilarity between the two sets by the dissimilarity between distribution of clusters.

Here we perform K-means clustering with 4 clusters on the combined set of reference locations and test locations and subsequently compare the cluster distribution. See Figure 7 for histograms of cluster distribution in reference samples and test samples and 8 for an overlay of clusters on the 200x200 grid corresponding to the area of the park. By inspection it is easy to see that the cluster distribution is significantly different, with the majority of elements in the reference set being placed in a single cluster, whilst the distribution of the test samples is much more uniform. This is explained by Figures 6 and 8 as we can see that the reference recordings are mostly taken in a relatively small area in the north-east quadrant, whereas a large proportion of the test recordings have been taken in the west half of the park.

This set of data could be investigated further and using alternative machine learning techniques such as mixture of Gaussian models however we limit the techniques applied in order to avoid increasing complexity.

Another method to assess the dissimilarity between the distributions of two sets of observations (here the reference recordings and Kasios test recordings) is to use a Gaussian Mixed Model approach.

Furthermore we want to look at the impact of the dump site on RCBP patterns of life. This is relatively straightforward as in order to do that we investigate the change in distance between recorded RCBP locations and the dump site. As the data is volatile it is justifiable to perform basic denoising

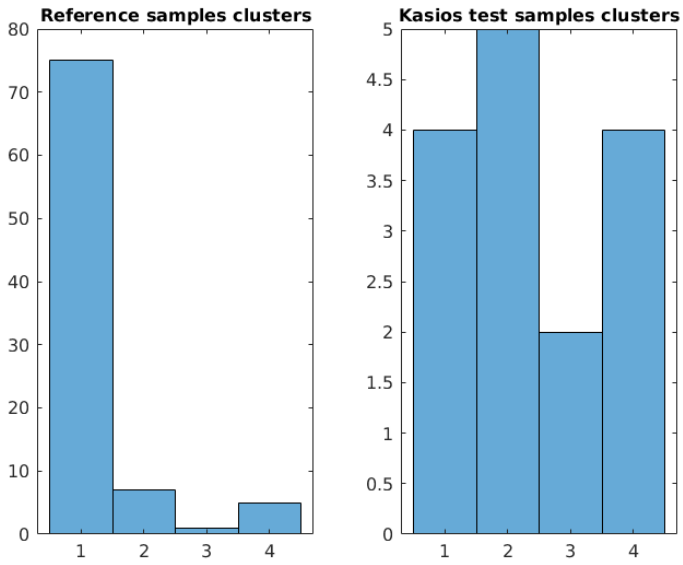

Figure 7: K-means cluster histogram

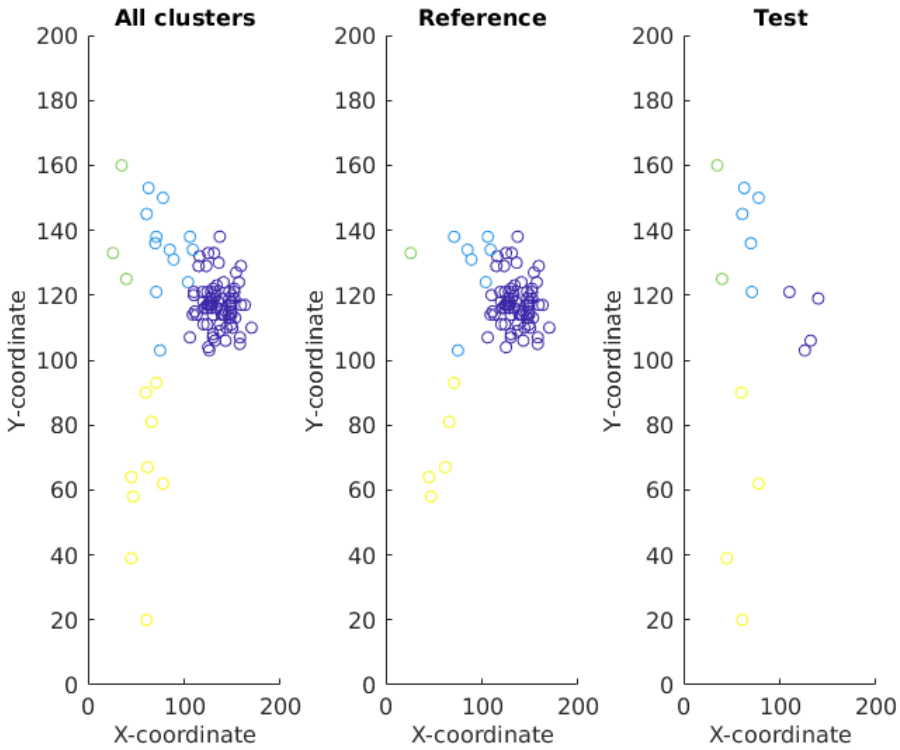

Figure 8: K-means cluster locations

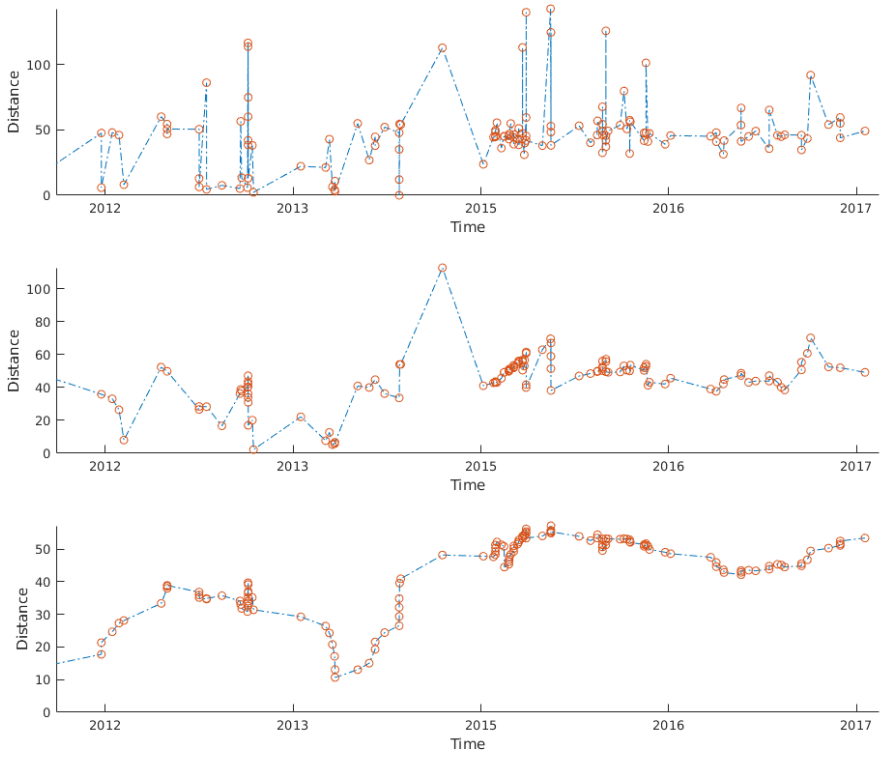

Figure 9: Distance from dump site: unfiltered(top), time-based moving average only (middle), fully smoothed (bottom)

(smoothing) prior to further analysis. This is done through time-based moving average with a 2-month window followed by ordinary moving average to remove outliers observed in low activity periods (see fig).

Note that we do not know the exact timescale from which the toxic waste dumping could have affected RCBP - we however do know that the process began in early 2015.

This shows that while RCBP did move away from the dumping site, this migration occured before the dumping process began. As such we can infer little to no information on the impact of the dumping site on RCBP patterns of life.

Finally we need to answer questions regarding trends in RCBP population - namely whether or not the population is decreasing and whether the dumping site location affects RCBP behaviour. Here we simply analyse the population in a manner very similar to the previous case.

To generate the timeseries we look at the number of RCBP sightings in every week from 2013 as a proportion of the total number of bird sightings in that period. To remove seasonal fluctuations we compute 1-year lagging running average. This can be seen in Figure 10. However we need to take into account both the lag associated with the moving average filtering as well as with the timescale required for the dumping process to affect population.

To help the analysis and make it easier to observe the timescale of changes we additionally generate a similar timeseries but filtered using a shorter time window. In addition it is appropriate to *zoom out* to see past trends in RCBP population. This is shown in Figure 11 Interestingly we can see a peak in RCBP activity just under a year after the dumping began. A general trend of increasing population

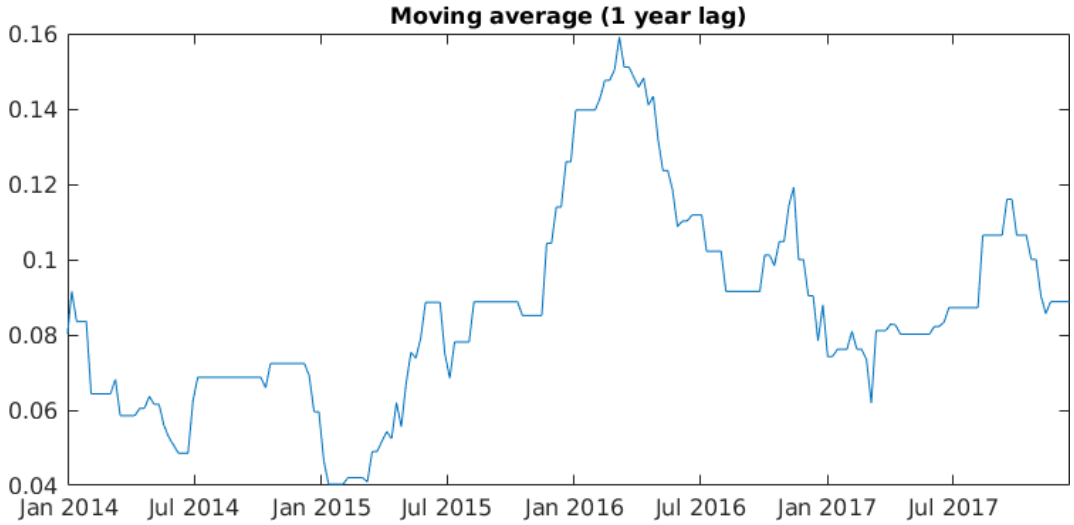

Figure 10: RCBP population changes with one year lag (as proportion of overall bird sightings)

until 2016 can be seen, but it is worthy of note that there have been periods of low activity in the past, which suggests some form of cyclical behaviour.

This data suggests that there has been a decrease in RCBP population, but the uncertainty regarding timescales makes it hard to argue that this is directly associated with Kasios misbehaviour. In fact given past cycles in RCBP behaviour it is possible to make a (weak) case that RCBP population is not, in fact, dwindling but what we are observing are simply cyclical variations.

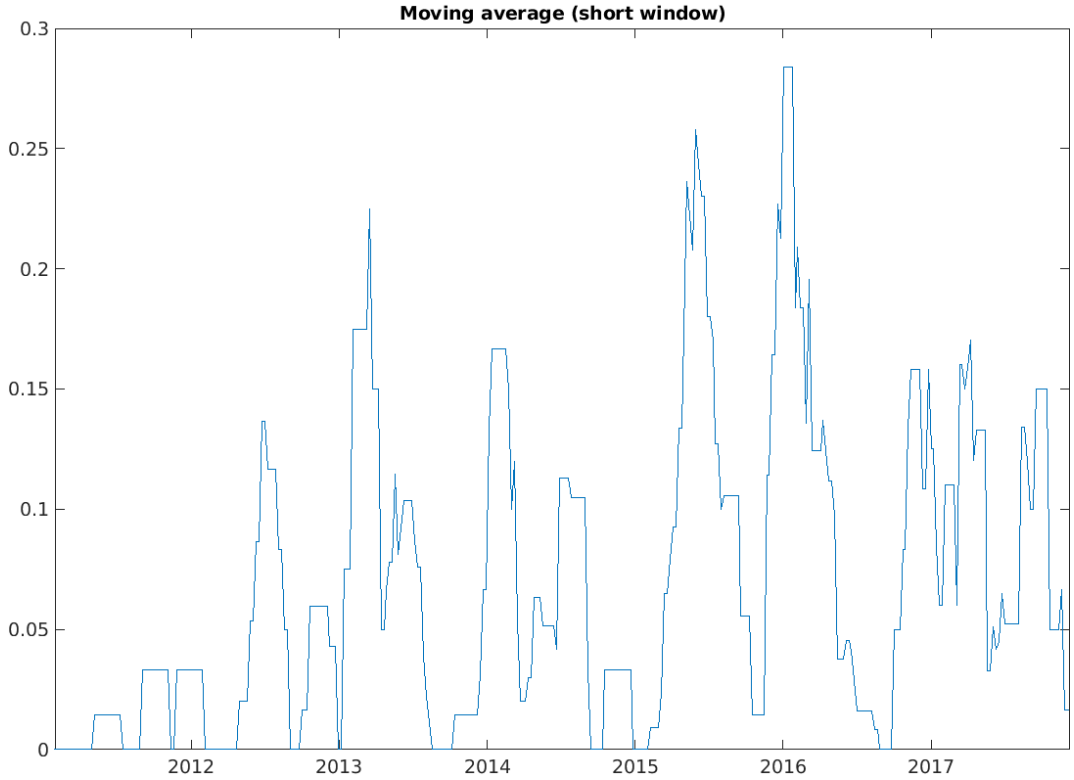

Figure 11: RCBP population changes with one year lag (as proportion of overall bird sightings)

3 Uncertainty modelling

In this section we discuss the methods used to generate mass assignments under theory of evidence from the different sources of information discussed throughout the previous section and how these pieces of evidence fit into the model described in subsection 1.2.

Note that this section is concerned more with providing a proof of concept / case study rather than identifying the *best* method of producing a basic belief assignment (bba) from data. In some places strict mathematical formalism is avoided in favour of guesswork in order to simplify and speed up the process.

3.1 Recording analysis

The data analysed in this section corresponds to the variable R in the situation model, concerned with whether or not the recordings provided by Kasios are genuine. We can model this in the belief function framework by representing the deep learning results obtained in the previous section as Dempsterian basic belief assignments.

As mentioned earlier, the methods used in this section to construct basic belief assignments are by no means *the* optimal methods to model uncertainty in this context - the purpose here is to demonstrate how such bbas can be obtained, rather than propose a state-of-the-art method of doing so.

Essentially two simple natural methods of bba generation exist - by estimating a Bayesian probability and correcting it given a certain degree of confidence or by estimating upper and lower bounds on probabilities. An obvious issue with these two approaches is that in larger frames of discernment it is only possible to directly populate the singleton sets and the universal set. This however is not a problem in this case study, as the variables we are concerned with can only take two values and as

such the power set will only consist of the singletons and the universal set.

From the classification task the trivial method of obtaining bba is to take the identification rate as our Bayesian probability and either the mean per class accuracy or RCBP accuracy as an indicator of our confidence levels.

Alternatively it is possible to estimate lower and upper bounds on probabilities using binomial theorem - although still relatively simple, this method is more complex whilst not necessarily being an improvement. A major challenge regarding bba generation given no ground truth is that the inherent uncertainty makes it impossible to directly compare different generation methods.

For the purpose of this case study we will seek to obtain least informative bba's (i.e. be as pessimistic as possible regarding our trust in the model), hence using mean per class accuracy as our confidence indicator.

Run	RCBP Id	MPC accuracy	R	$\bar{R}$	$\{R, \bar{R}\}$
1	0.0667	0.3567	0.02378	0.33292	0.6433
2	0.1333	0.4039	0.0539	0.350	0.5961

Table 2: Obtained basic belief assignments for $\Theta_R = \{R, \bar{R}\}$

Furthermore we need to obtain a basic belief assignment from the autoencoder results. The approach used here is based on estimating upper and lower limits on probability. Recall that the reconstruction error can be treated as an estimate of how "atypical" an element is. A significantly larger reconstruction error for the test cases would imply that they are in some way anomalous.

We estimate the probabilities as follows. In order to obtain the upper limit on probability (the plausibility value) we model the reference recordings reconstruction error distribution as a normal distribution and compute the probability that 15 samples (the number of test recordings) have a mean equal or higher to the test mean. This yields the plausibility value of $Pl(R) = 0.4364$.

We can estimate the lower limit on probability of R by looking at the proportion of test recordings where the reconstruction error is lower than reference mean. This implies $Bel(R) = 0.2636$.

Finally we discount the belief function using a confidence value of 0.8. This yields the following basic belief assignment

R	$\bar{R}$	$\{R, \bar{R}\}$
0.2109	0.4509	0.3382

Table 3: Obtained basic belief assignments for $\Theta_R = \{R, \bar{R}\}$

3.2 Test files metadata

In this section we attempt to generate a basic belief assignment from the analysis of Kasios files location metadata.

Again we use a very simple method to obtain a bba from the k-means analysis, by first analysing the reference results. It is clear that there exists a *typical* cluster and three *atypical* cluster with varying degrees of rarity. Hence for a test sample, we could argue that membership in the typical cluster provides support for L and membership in the atypical cluster provides support for $\bar{L}$.

The next step is to estimate the degree to which cluster membership supports the corresponding hypothesis. For simplicity let us assume that we can derive it directly from the probability distribution of clusters in reference data.

Furthermore to make the results more conservative and reduce the impact of selection of $k = 4$ for clustering let us treat the three atypical clusters as a single one.

As such each of the clusters provides support for the appropriate hypothesis to the degree of 0.8523 (the reference cluster distribution is $[75, 7, 1, 5]$ yielding a frequentist probability distribution of $[0.8523, 0.0795, 0.0114, 0.0568]$.

The final step is bba discounting - here we use a conservative value of 0.8. The final bba is shown in table 4.

L	L	$\{L, \bar{L}\}$
0.2131	0.4687	0.3182

Table 4: Obtained basic belief assignment for $\Theta_L = \{L, \bar{L}\}$

3.3 Changes in RCBP patterns of life

This corresponds to D and P variables in the model.

As stated in the previous section it isn't possible to assess the direct impact of the dumping on RCBP patterns of life hence we assume $m_{\Theta_D}(D, \bar{D}) = 1$. However we also need to bear in mind our prior assumption that methylosmene dumping *has* an impact on RCBP population.

Therefore let us assume the following overall belief in D

P	$\{P, P\}$
0.7	0.3

Table 5: Prior basic belief assignment for $\Theta_D = \{D, \bar{D}\}$

There are two sources describing the changes in RCBP population. One is a prior belief (directly from the 2017 and 2018 challenge description). Let us assume this to be 80% reliable.

P	$\{P, \bar{P}\}$
0.8	0.2

Table 6: Prior basic belief assignment for $\Theta_P = \{P, \bar{P}\}$

In addition as per discussion in the previous section we can estimate our belief in $\{P, \bar{P}\}$. As we lack formal mathematical framework, we need to intuitively combine: weak evidence that the population is not dwindling (cyclical patterns), relatively strong evidence that the population is falling and our confidence in this assesment (relatively low, due to conflicting beliefs, lack of information regarding timescales and lack of formal methodology). As such we propose the following bba

P	P	$\{P, P\}$
0.4	0.1	0.5

Table 7: Basic belief assignment for $\Theta_P = \{P, \bar{P}\}$ obtained from analysis of RCBP patterns of life

4 Evidence combination

The evidence described in previous sections now needs to be combined. As this is computationally complex but intuitively straightforward it is not covered in much detail here but rather only key steps are explained.

Figure 12 shows the situation assesment model annotated using the belief masses for the distinct variables after the combination process. Masses m_R and m_P are obtained by combining source masses (in case of m_R) or source masses and a mass derived from m_K (m_P) using Dempster's rule of combination.

Some combination processes are less straightforward. In order to obtain m_K both m_L and m_R are vacuously extended so that they are both defined on the new frame of discernment $\Theta_{LR} = \{LR, \bar{L}R, L\bar{R}, \bar{L}\bar{R}\}$ and subsequently combined using Dempster's rule $m_{LR}^{\Theta_{LR}} = m_L^{\Theta_{LR}} \oplus m_R^{\Theta_{LR}}$. Following that the mass is transformed so that

$$\begin{aligned} m_K(K) &= m_{LR}^{\Theta_{LR}}(LR) \\ m_K(\bar{K}) &= \sum_{LR \cap \theta = \emptyset} m_{LR}^{\Theta_{LR}}(\theta) \\ m_K(K, \bar{K}) &= \sum_{LR \cap \theta \neq \emptyset, \theta \neq LR} m_{LR}^{\Theta_{LR}}(\theta) \end{aligned}$$

A similar but simpler process is used to obtain an intermediate mass $m_K^{\Theta_P}$, one of the belief assignments contributing to m_P. This transformation is based off the logical assumption that $K \rightarrow \bar{P}$

$$\begin{aligned} m_K^{\Theta_P}(\bar{P}) &= m_K(K) \\ m_K^{\Theta_P}(P, \bar{P}) &= 1 - m_K(K) \end{aligned}$$

The *weak implication* $\bar{K} \overset{\delta}{\dashrightarrow} F$ is used to obtain the intermediate belief function $m_K^{\Theta_F}$ as follows

$$\begin{aligned} m_K^{\Theta_F}(F) &= \delta \times m_K(\bar{K}) \\ m_K^{\Theta_F}(F, \bar{F}) &= 1 - \delta \times m_K(\bar{K}) \end{aligned}$$

An important addition to the situation chart is th new node I which represents the assumption of innocence. Note that from the situation model all the variables feeding into F can only provide evidence that Kasios is faulty or no evidence at all (ignorance). The node I represents the prior belief that Kasios is indeed not at fault.

From the above and with the values $I = 0.6$ and $\delta = 0.4$ we obtain the following overall bba on $\Theta_F = \{F, \bar{F}\}$

$$\begin{aligned} m_F(F) &= 0.5345 \\ m_F(\bar{F}) &= 0.2793 \\ m_F(F, \bar{F}) &= 0.1862 \end{aligned}$$

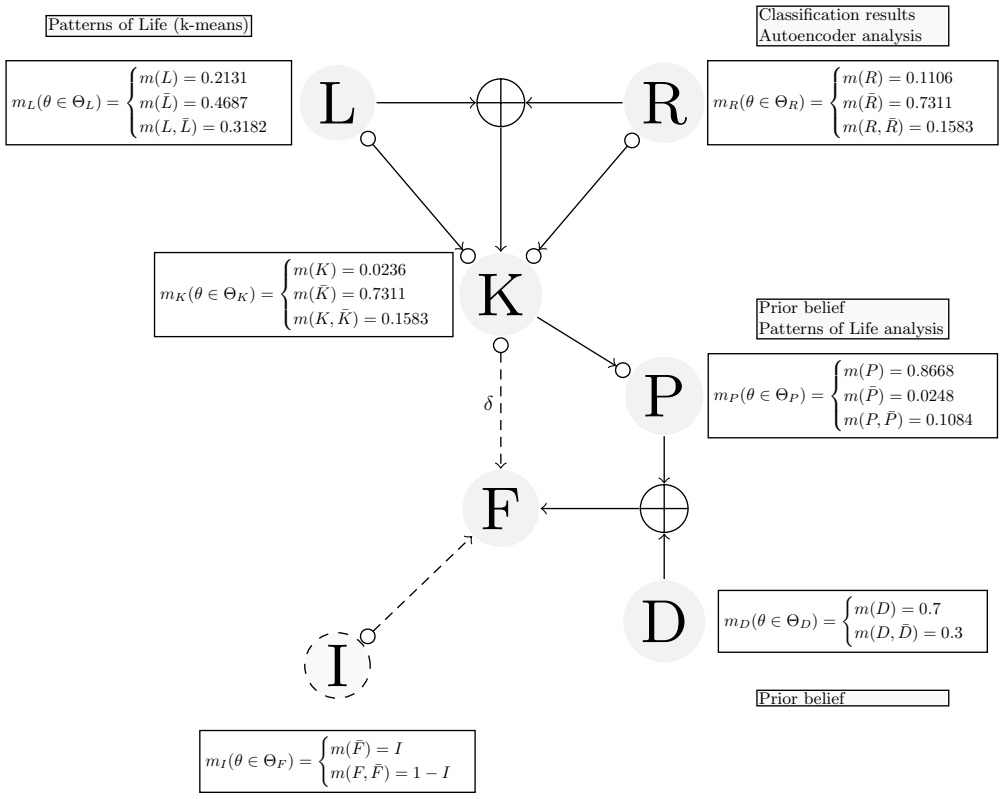

Figure 12: Graphical situation model annotated with belief masses and contributing sources

5 Decision making and provenance analysis

We simplify the decision making process to identification of the most likely outcome in the set $\Theta_F = \{F, \bar{F}\}$. Using any classical decision making method we obtain $\hat{\theta} = F$.

Now let us consider the sources which drive this decision. Figure 13 shows the provenance scores for the variety of information sources contributing to F. Note that I is missing from the graph as we know it only contributes towards $\bar{F}$.

The provenance values are computed using the provenance measure proposed in [ref1] and [ref2].

This shows us, unsurprisingly, that our belief in RCBP population falling and the negative effect of methylosmene is the key piece of evidence in judging Kasios responsible. The recordings provided by Kasios were deemed untrustworthy and hence had a minimum effect on our beliefs in RCBP population trends. PoL analysis didn't have a significant impact and hence the key piece of evidence driving this entire system was our initial belief in dwindling RCBP population. This makes intuitive sense, as the initial premise was that this population is indeed falling, and the files provided by Kasios were supposed to act as counterevidence, which they failed.

We could argue however that making a decision on K was the key aspect of this entire problem, as the pattern of life analyses were insignificant compared to the prior. However had our evidence that $\theta_K = \bar{K}$ been weaker, it could have significantly affected the outcome. Hence a similar analysis is performed for K with its results shown in Figure 14. It can be seen that the key contributor are the classification results, which is again unsurprising.

Finally we want to investigate the impact of δ and I, two assumptions we have made which may

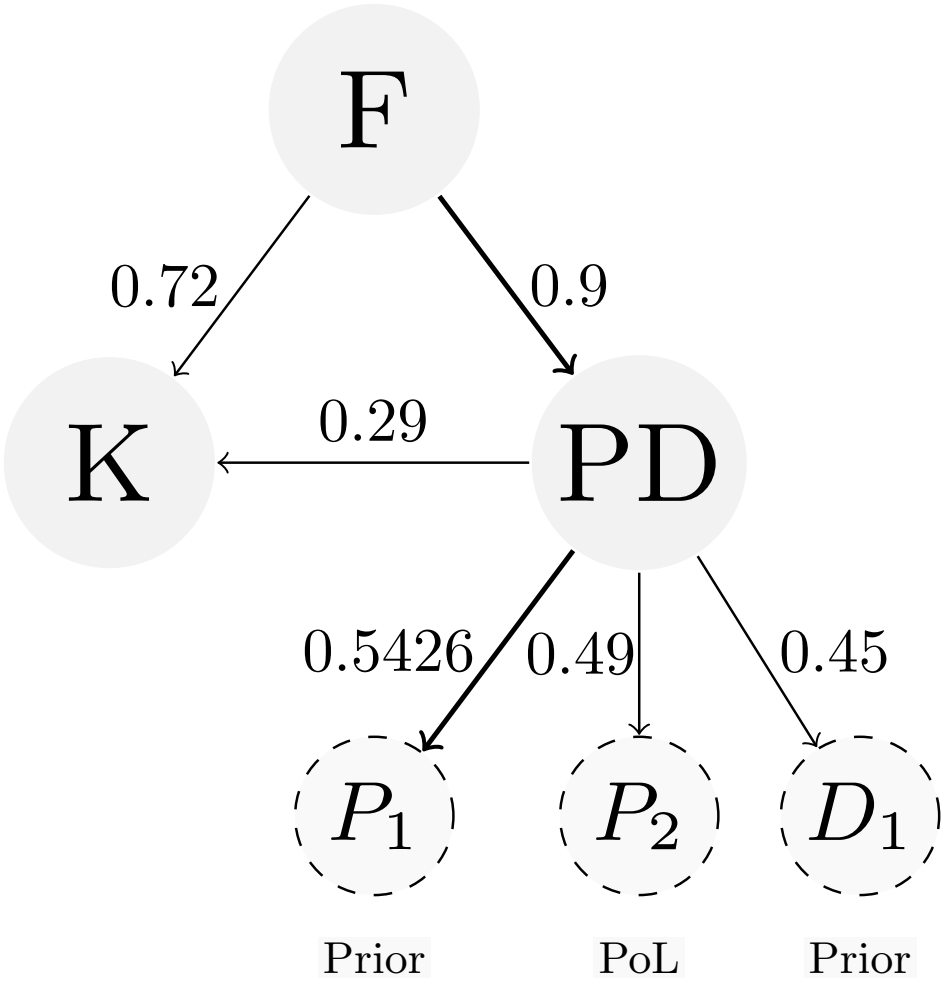

Figure 13: Provenance values on θ_F given decision $\theta_F = F$

have a significant impact on the decision. On Figure 15 the relative impact of these two variables is shown. Below the green line, the decision $\hat{\theta_F} = F$ is made, whereas the opposite happens above.

It can be seen that given our choice of $I = 0.6$ the choice of δ is irrelevant. In fact the value of I could be increased up to 0.75 before the decision would change, showing relative robustness of our results.

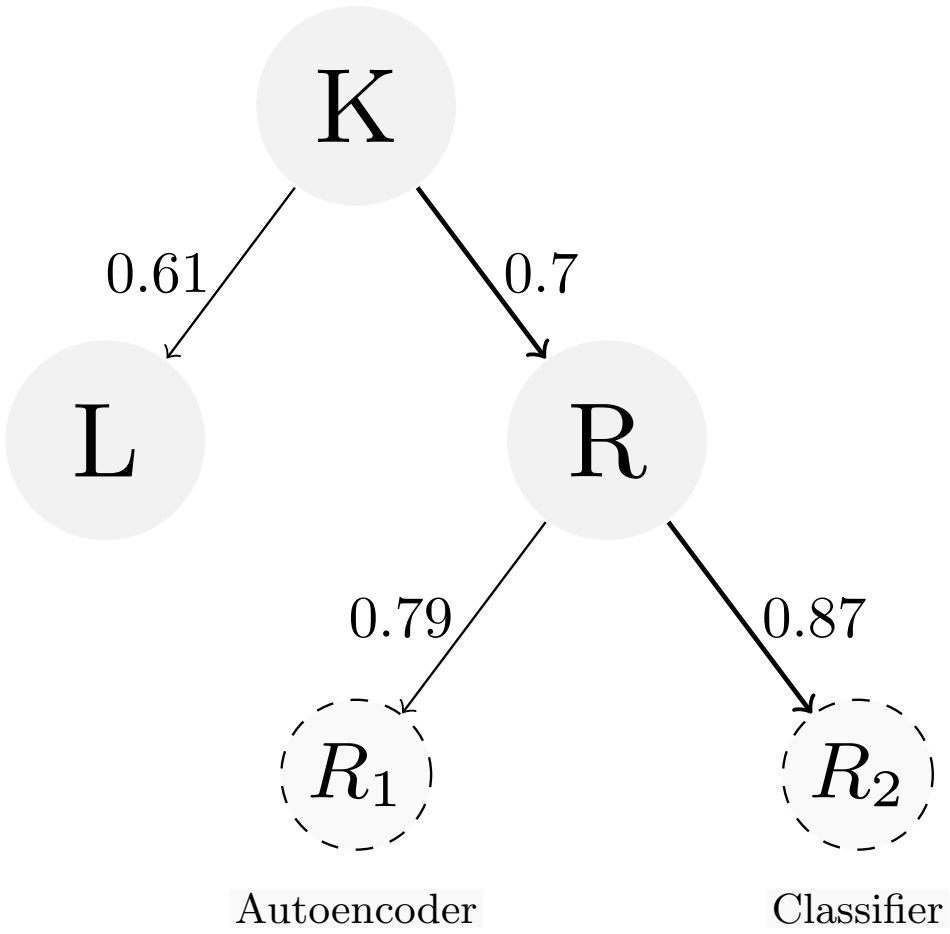

Figure 14: Provenance values on θ_K given decision $\theta_F = \bar{K}$

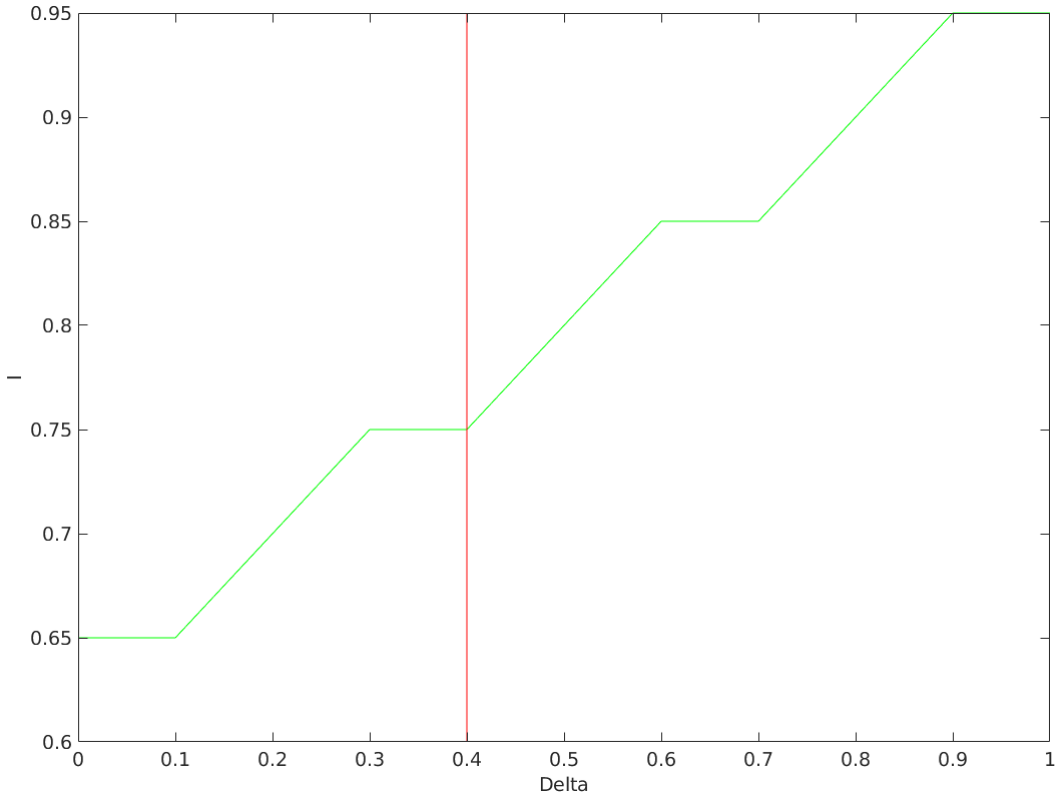

Figure 15: Decision made $\hat{\theta_F} = F$ and $\hat{\theta_F} = \bar{F}$ as values for I and δ are varied