Inferring the Most Probable Maps of Underground Utilities using Bayesian Mapping Model

Muhammad Bilal[1,*], Jennifer Muggleton[3], Emiliano Rustighi[3], Hugo Jenks[4], Steve R. Pennock[4], Phil R. Atkins[5] and Anthony Cohn[2]

[1]University of California, Center for Environmental Implications of Nanotechnology (UCCEIN), University of California, Los Angeles (UCLA)
Los Angeles, USA

[2]School of Computing, University of Leeds, Leeds, United Kingdom

[3]Institute of Sound and Vibration Research, University of Southampton, Southampton, UK

[4]School of Electronic and Electric Engineering, University of Bath, Bath, UK

[5]School of Electronic, Electrical and Computing Engineering, University of Birmingham, Birmingham, UK

Email: m.bilal@ucla.edu[*]. * - Corresponding author.

Abstract

Mapping the Underworld (MTU), a major initiative in the UK, is focused on addressing social, environmental and economic consequences raised from the inability to locate buried underground utilities (such as pipes and cables) by developing multi-sensor mobile device. The aim of MTU device is to locate different types of buried assets in real time with the use of automated data processing techniques and statutory records. The statutory records, even though typically being inaccurate and incomplete, provide useful prior information on what is buried under the ground and where. However, the integration of information from multiple sensors (raw data) with these qualitative maps and their visualization is challenging and requires the implementation of robust machine learning/data fusion approaches. An approach for automated creation of revised maps was developed as a Bayesian Mapping model in this paper by integrating the knowledge extracted from sensors raw data and available statutory records. The combination of statutory records with the hypotheses from sensors was for initial estimation of what might be found underground and roughly where. The maps were (re)constructed using automated image segmentation techniques for hypotheses extraction and Bayesian classification techniques for segment-manhole connections. The model consisting of image segmentation algorithm and various Bayesian classification techniques (segment recognition and expectation maximization (EM)) provided robust performance on various simulated as well as real sites in terms of predicting linear/non-linear segments and constructing refined 2D/3D maps.

Keywords: *MTU Sensors, Most probable maps, Bayesian Data Fusion, Image processing, Bayesian Regression.*

1. Introduction

The costs associated with street works in the UK is of critical consideration due to the vast majority of utilities buried underneath the roads and their repair/(re)installation (£7b annual) (McMahon, 2005). The types of utilities buried under the ground are diverse and their amount is notoriously large which makes excavation a challenging task in order to upgrade these underground networks. In addition, the statutory records of underground networks are typically incomplete and inaccurate particularly for old street works (M Burtwell, 2004). An important undertaking is to develop schemes to detect what is buried underground that could be associated to their records and could become cost saviour. A multi-sensor mobile laboratory MTU (Underworld, 2011) was developed which consists of multiple sensors capable of deploying several approaches to detect different types of infrastructure. The multi-sensor device, was designed to assess the feasibility of a range of potential technologies that can be combined into a single device to accurately locate buried pipes and cables. The potential technologies included ground penetrating radar (GPR), low-frequency quasi-static electromagnetic fields (LFEM), passive magnetic fields (PMF) and low frequency vibro-acoustics (VA) and significant advances have already been made (Royal et al., 2011, Royal ACD, 2010).

The location techniques combined by MTU provide significant advantages over other commercially available techniques (Ashdown, 2000) for detecting wide variety of utilities and control trials were taken for few commercial sites. As a result, excavations necessary for maintenance and repair can be largely reduced using such a device equipped with a variety of sensor technologies. An important undertaking is to use heterogeneous information from these sensors and build refined maps of buried utilities in real time. However, due to the heterogeneity in features of utilities and ground properties, it is challenging to develop a general technique that

could assess heterogeneous information and handle the uncertainties associated to this task. The

integration of information obtained from multiple sensors on MTU is of critical importance in

order to make sense of the data before providing a precise information on a site. The knowledge

obtained from different sensors presents itself non-symbolically i.e. the delivered data is

essentially an image representing what the sensor "sees" underground. In contrast, utility records

are almost universally represented symbolically i.e. they are stored in a spatial database as records

with a vectorized representation of their spatial position, along with attribute information (such as

material, diameter). It is therefore challenging to provide a useful and accurate representation of

the data acquired from a variety of sensors. Therefore, a data fusion approach consisting of

automated techniques for data extraction and integration was imperative.

Data fusion includes fusing raw data that can be used to automatically extract symbolic

hypotheses about the presence of a buried asset. Not all these detections may actually correspond

to utility assets, and equally there may be assets which are not detected by the sensor (because the

sensor cannot detect that kind of asset or the environmental conditions are unfriendly – such as

saturated clay for GPR). We associate a probability distribution with every such hypothesis

reflecting possible errors in the measurements and hypothesis extraction process. These

geographical positions (x, y) and depths (z) were used as input to the next stage of the mapping

system. A variety of Artificial Intelligence (AI) techniques and algorithms were implemented such

as Bayesian Data Fusion (BDF), image segmentation, orthogonal distance hyperbolic fitting, and

weighted variation. The algorithms for automated data processing and map (re)construction were

developed for real time operative capability of MTU device.

The map (re)construction model developed in this work was an improvement over (Chen and

Cohn, 2011) which was initially designed only for 2D construction of the map assuming that it

consists of only linear segments. In addition, the data preprocessing for hypotheses extraction in (Chen and Cohn, 2011) was not combined as a complete model and it was assumed that the hypotheses were extracted from GPR images using an iterative clustering/classification techniques prior to data fusion tasks. Simple clustering/classification algorithms for hypotheses extraction such as k-means or Dbscan were restricted in several ways for asset classification problem when developing real time maps. For example, traditional k-means clustering algorithm creates the clusters based on Euclidean distance of each data point to the centroids (initially selected randomly). Also the number of clusters to be created is known in k-means algorithm. Depending on statutory records to identify the number of segments was not reliable as, even providing valuable information, they are inaccurate and may contain incomplete information. Dbscan (Sander et al., 1998) also separates the clusters based on Euclidean distance without providing the desired number of clusters to be generated as prior. However, Dbscan requires radius in order to differentiate the clusters that is used as a criterion for decision making on number of clusters. The Euclidean distance between parameters is important in both approaches which is helpful in situations where clustering is only distance based.

Neira (Neira, 2001) developed a data association model for addressing the problem of robust data association for simultaneous vehicle localization and map building which was an improvement over gated nearest neighbor (NN) (Bar-Shalom, 1987) for tracking problems that successfully rejects spurious matching and provides optimal solutions in terms of pairs of matching in cluttered environments. The correlation between measurement prediction errors in 2D space in cluttered environment provides robust data association with an efficient traversal of the solution space. However, the directional errors (linearity) caused mismatching of the segments with manholes using the hypotheses extracted from the sensors. Abhir and Roland (Abhir Bhalerao

92   2001) also used a Multi-resolution Fourier Transform (MFT) for capturing sufficient shape and

93   orientation of objects within a given image. The use of statistical analysis and camera projections

94   to estimate the location/orientations of line segments in 3D image was also implemented for

95   similar linear segment construction problems (Dong-Min and Dong-Chul, 2009, Ting-Wang and

96   Qing, 2010). However, these approaches are only limited to an image of objects and segments

97   which is used to reconstruct a 3D image. MTU mapping, on the other hand, is multi-source data

98   fusion approach to integrate information from multiple sources and produce most probable maps

99   utilizing advanced machine learning/data mining techniques. For linear segment fitting, significant

100   amount of literature report the use of different regression models including EM algorithm that can

101   efficiently fit at higher accuracy levels (Ward et al., Ester et al., 1996, Sanquer et al., 2011,

102   Delicado and Smrekar, Werman and Keren, 1999, Friedman and Popescu, 2004). The

103   classification of data samples based on its source as distinguished by MTU sensors is, however,

104   lacking in these approaches as these algorithms were developed for regression scenarios. In

105   addition, the connection establishment (manhole-segment) was not considered as an underlying

106   issue as only the general regression was covered.

107   The Bayesian mapping model is capable of using automated techniques for hypotheses

108   extraction, classification, segment recognition and connection establishment with the associated

109   manholes. The model workflow was divided into three major components which were (1)

110   hypotheses extraction from sensor input images using variety of image segmentation and Bayesian

111   clustering techniques, (2) segment (pipe or cable section detected by the sensor) recognition and

112   noise removal techniques using the extracted hypotheses, and (3) segment-manhole connection

113   using EM algorithm. A complete use case can be tested using real time mapping model where

114   hypotheses extraction techniques were combined with iterative connection establishment and

115 visualization techniques. Several simulated as well as real sites were tested and it was

116 demonstrated that the model is robust in various conditions where statutory records were

117 unavailable and the sensor readings were sparse. The segments were recognized and noise was

118 removed successfully in various situations for mapping the utilities demonstrating the ability of

119 model to work in real time complex situations.

2. Materials and Methods

121 The model for Bayesian mapping followed the workflow depicted in **Fig 1**. The sequential steps

122 in model workflow were as follows; (1) data preprocessing, (2) segment recognition, and (3)

123 segment-manhole connection. The datasets consisted of both qualitative (raw images from MTU

124 sensor device, statutory records (utility maps)) as well as quantitative information (manhole

125 surveys providing information on directions and depths). The data from these sources were

126 integrated together in the form of positions (x, y), depths (z) and directions (Θ) as an input for

127 segment recognition algorithm. Manhole surveys were helpful in determining the types of buried

128 utilities at a site and estimating their directions and depths. Bayesian recursive algorithm as known

129 as EM was developed for segment-manhole connection. EM is an iterative algorithm for finding

130 the maximum-likelihood estimation of a set of parameters for a specific distribution in a statistical

131 model when it contains unobserved latent variables (Do and Batzoglou, 2008). Based on partially

132 connected information from different sources and segments recognized from segment recognition,

133 EM algorithm was used to identify most probable connections of segments with assets and connect

134 them in an iterative manner. The depth information from manholes and sensor readings were then

135 used in conjunction with EM output to produce refined 2D/3D maps of an investigated site. The

136 model workflow steps are described as follows:

6
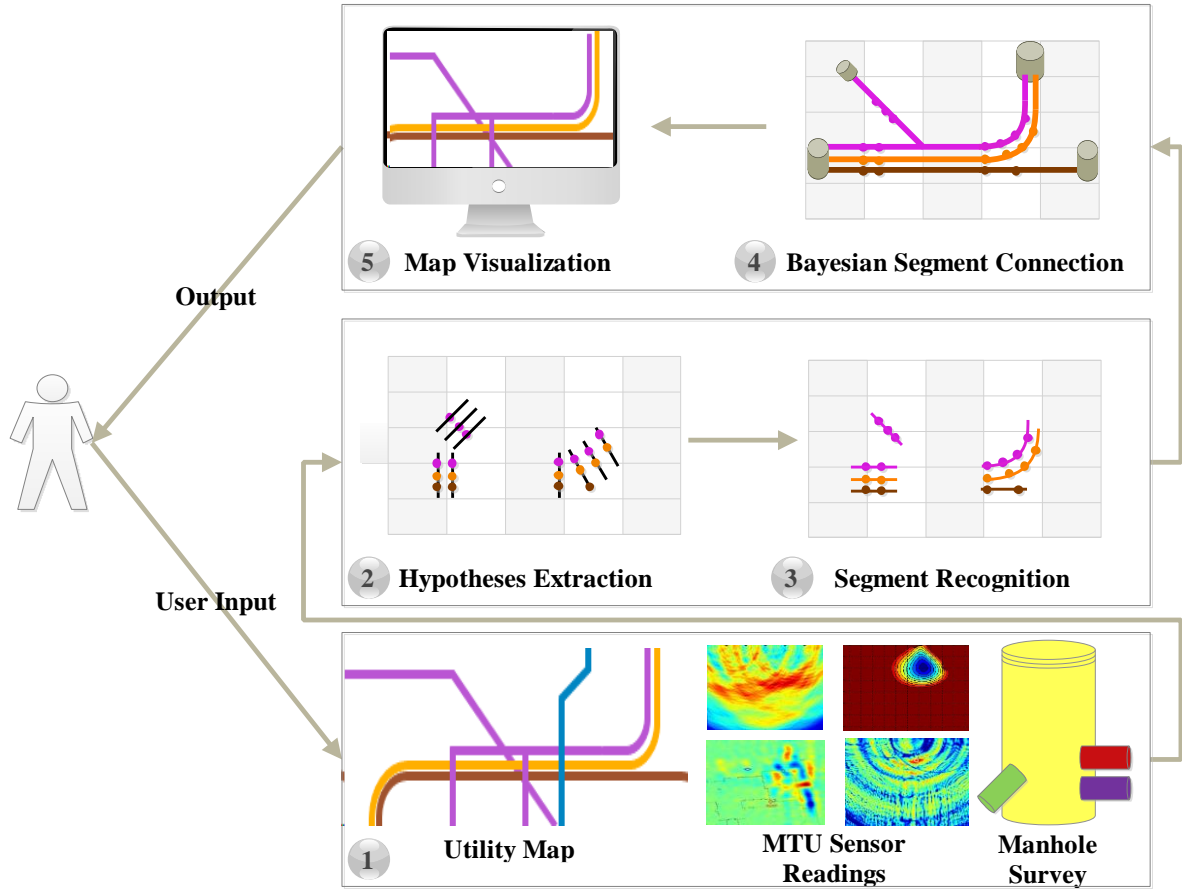
Figure 1: Model Workflow (Data preprocessing for hypotheses extraction following segment recognition, Bayesian segment connection and refined map visualization)

2.1. Data Preprocessing: Data for model development were collected from 3 different sources which were;

2.1.1. Statutory Records: The utility/statutory records consisted of maps of buried assets as ground truth which provided information on approximate directions and positions of these assets. Since these records are inaccurate, their position information was not included in segment recognition algorithm. However, the information pertaining to the type of buried asset and their approximate directions was helpful in segment-asset connection algorithm. Therefore the positions and directions $(x, y, \Theta)^{\mathrm{T}}$ from these records were used for segment-manhole connection establishment.

146   **2.1.2. MTU Sensor Raw Data:** The raw data included heatmaps (**Fig. 2**) of investigated sites that

147   were produced from various MTU and commercial sensors. The sensors used in MTU mobile
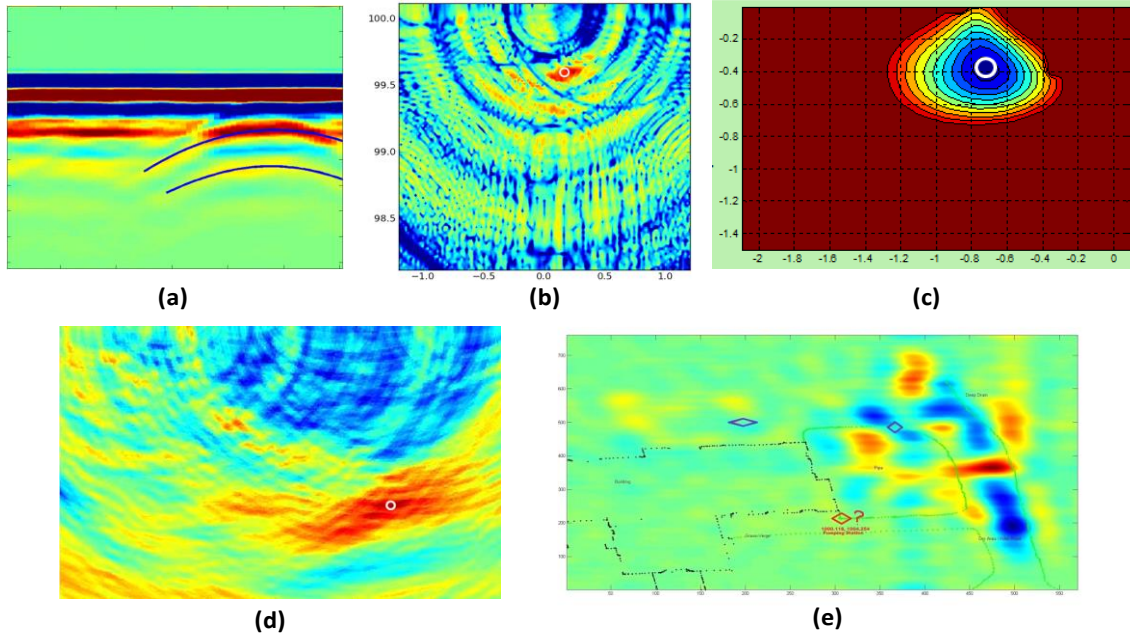
Figure 2: Images taken from MTU sensors and processed to mark the hypotheses (hyperbolae and circles). The images were created from the raw data collected from **(a)** Commercial GPR, **(b)** MTU GPR, **(c)** Passive Magnetic Field, **(d)** Vibro acoustics, and **(e)** Low Frequency Magnetic Field.

148   apparatus include Ground Penetrating Radar (GPR), Passive Magnetic Field (PMF), Vibro

149   Acoustic, and Low Frequency Electro-Magnetic (LFEM). Briefly, the sensors and their

150   functionality are explained in this section. Detailed information on each sensor can be found

151   elsewhere (Dutta et al. 2013; Muggleton et al. 2011; Thomas et al. 2009).

152   **2.1.2.1. Ground Penetrating Radar (GPR):** GPR locates buried utilities (metallic and non-

153   metallic) by transmitting electromagnetic waves into the ground and collecting the response

154   (waves reflected from the objects underground). A GPR scan produces an image as a collection of

155   multiple A-scans of reflected waves received at different wave-travel times and integrating them

156 into a B-scan. In B-scan, the rows represent the depth (by utilizing the reflected waves) and the

157 columns represent the horizontal positions of the scan lines.

158 **2.1.2.2. Passive Magnetic Field (PMF):** PMF is useful for detecting the cables by utilizing the

159 magnetic field generated by current in a cable as well as neighboring objects which generate

160 magnetic field (may also be from the current flow leak from a buried cable). The cluster centroids

161 (**Fig. 2c**) are multiplied by the surveyed map dimensions ($m \times n$ image) which are then divided by

162 surveyed lengths and depths, respectively, to calculate the (x, y, z) coordinates on the map.

163 **2.1.2.3. Vibro Acoustic (VA) Sensors:** VA device at the MTU uses seven geophones which excite

164 the buried utility at survey location and receive raw response on the surface which is then utilized

165 to generate the cross sectional image via frequency domain transformation. The areas of cross

166 sectional image representing maximum intersections of the responses from seven geophones are

167 extracted as hypotheses (**Fig. 2d**).

168 **2.1.2.4. Low Frequency Electromagnetic Field (LFEM):** LFEM approach is based on injecting

169 the current into the ground and measuring the voltage on the surface via coupled plates moved

170 along the surface. The LFEM approach at MTU is operated in a targeted grid location and the

171 resulting image represents the underground structure (**Fig. 2e**).

172 Using the above approaches, qualitative raw images consisting of $(x, y, z)^{\mathrm{T}}$ information (x-axis

173 as the width of sensor device and y-axis as the depth of the investigated surface (typically 2-4

174 meters deep)) were collected and processed to extract hypotheses which were utilized for 2D/3D

175 map reconstruction. The orientation of MTU sensor device provides approximate direction of

176 buried asset as the survey is usually taken in the direction perpendicular to the orientation of asset.

177 **2.1.3. Manhole Survey:** Manhole surveys for real sites were conducted to collect supporting

178 information on estimated directions and depths of buried assets. In addition, types of buried assets

179   were also recorded in order to distinguish between them when constructing site maps. The

180   quantitative information from manhole surveys therefore included $(x, y, z, \Theta)^{\mathrm{T}}$ as well as the type

181   of buried asset as qualitative information.

182   **2.1.4. Hypotheses Extraction:** Hypotheses (processed coordinates of the location detected by

183   MTU sensors) extraction was implemented on raw sensor image data to automatically estimate the

184   positions from surveyed locations. The images were segmented into clusters of foreground

185   (hypotheses) and background pixels using unsupervised image segmentation techniques. The

186   centroids of clusters were identified as approximate positions of assets. Sensor output image is

187   often noisy and may not provide an accurate location. Additionally, multiple hypotheses locations

188   may be reported in close proximity when different types of assets (or a non-asset object) are

189   present. From multiple images of a single surveyed location, the most probable locations of assets

190   were estimated using Bayesian weighted average technique, i.e.

191     $h(x_i) \approx N(x_i, \mu, \Lambda)$ (1)

192   In the above equation, x_i denotes hypothesis i collected from the image, μ is the mean of selected

193   hypotheses, and Λ is covariance matrix of the hypotheses. Maximum A Posteriori (MAP) position

194   of an asset was obtained using $h(x) = \arg\max \prod_{i=1}^{n}(h(x_i))$ where n is the number of hypotheses.

195   Prior to segment recognition from groups of hypotheses, the identification of noise was

196   conducted in terms of points with no manhole connection within a given previously tested

197   threshold ($\delta = 2$ meters). Each hypothesis point was validated using: (1) point to line (nearest

198   buried utility in statutory record) distance and (2) manhole directions. The distance d_{jh} from each

199   pipe segment j to the point h was calculated using

$$d_{jh} = \frac{\left|(x_{je} - x_{js})(y_{js} - y_h) - (x_{js} - x_h)(y_{je} - y_{js})\right|}{\sqrt{(x_{je} - x_{js})^2 + (y_{je} - y_{js})^2}}$$

(2)

200   Where (x_{js}, y_{js}) and (x_{je}, y_{je}) are the start and end points of the pipe segment j respectively, and

201   (x_h, y_h) is the location of hypothesis point. The point h was considered as an orphan point if its

202   distance from a pipe segment j was greater than δ. i.e.

203   For all $k = 1,2,..., K$ manholes of same type as that of given pipe segment

204    $d_{hk1} = \sqrt{\left(x_{js} - x_k\right)^2 + \left(y_{js} - y_k\right)^2}$

205    $d_{hk2} = \sqrt{\left(x_{je} - x_k\right)^2 + \left(y_{je} - y_k\right)^2}$

206   Where (x_k, y_k) is the location of the manhole k. The orphan points were obtained using

207    $h = \begin{cases} orphan & if\ d_{hk1} > \delta \wedge d_{hk2} > \delta \\ h_r & otherwise \end{cases}$

208   And the probability that the point h belongs to a segment j was then calculated as

209    $S_j = \arg\max_j (P_{hj})$ where $P_{hj} = F(h \rightarrow j | d_{jh}) \approx N(h, d_{jh})$

210   **2.2. Segment Recognition:** The hypotheses obtained from above step were given as input to

211   segment recognition algorithm which estimated the segments by joining hypotheses together based

212   on their positions and direction. The input to segment identification algorithm included the

213   hypotheses and the directions in which the surveys were taken (usually from left-to-right). The

214   linear segments were iteratively classified from groups of hypotheses which were located within

215   the range of an angle of up to $\varepsilon = \pm 0.08$ (in radians). The algorithm starts with first two chosen

216   hypotheses as assigned to first class (starts with class=1), and the directions from first hypothesis

217   to all other hypotheses are calculated. The next point is chosen and assigned to the same class if

218   their difference in directions is $<= \varepsilon$ otherwise the chosen point is assigned a new class. The

11

219   algorithm is repeated for each hypothesis and provides the output as points assigned to a class

220   which are then joined using least squares, circle and/or polynomial fitting techniques. The

221   workflow of the algorithm is given in following section.

222   Input: $H = (x, y)^T$ where x and y are the spatial locations of a set N of hypotheses in the map

---

Algorithm: Segment recognition algorithm workflow

A: $n \times 2$ *matrix of* $arctan(m_i)$ *and* C_i *where* m_i = *slope of* i^{th} *from first point and* C_i= *class*
   Initialize $C_1 = 1$, $rad = 0.08$, *and find* $arctan(m_i)$ *for* $i \in 2, ..., n$
   WHILE: *all hypotheses points in H are not visited and not assigned a class C*
   IF: $|arctan(m_i) - arctan(m_1)| > rad$
       FOR: $j \in 1\ to\ i$
           IF: $|arctan(m_i) - arctan(m_j)| < rad$ *% if closer to another existing class*
               Assign existing class C_j
           ELSE:
               Assign new class $C_{total+1}$
       END
   ELSE:
       FOR: $j \in 1\ to\ i$
           IF: $|arctan(m_i) - arctan(m_j)| < rad$ *% if closer to another existing class*
               Assign existing class C_j
           ELSE:
               Assign new class $C_{total+1}$
       END
   *Classify points of same class into separate clusters*

END

223

224   To separate the clusters of linear segments from curved segments, the goodness of fit (R^2) was

225   checked and compared to threshold $\varepsilon_1 \geq 0.99$ (0.99 confidence). Each cluster with $R^2 \geq 0.99$ was

226   separated as a linear pipe segment and the remaining clusters were considered as curved or circular

227   segments. For each hypothesis point in curvature analysis, the tip of the vector $r(t) =< x(t), y(t) >$

228   traces out a path in the plane where t represented MTU survey line and y represented the change

229   in positions represented by hypotheses. The relative positions of the hypotheses points were

230   therefore represented as follows $r'(t) =< x'(t), y'(t) >$ where $x'(t) = \dfrac{\partial x}{\partial t}$ and $y'(t) = \dfrac{\partial y}{\partial t}$. The unit

231   tangent vector was calculated as $T = \dfrac{r'(t)}{\|r'(t)\|}$ where $\|r'(t)\| = \sqrt{x'(t)^2 + y'(t)^2}$. In order to determine the

232   variation in the position (y) with respect to (x), the difference interval was obtained using

233    $\eta = \left\| \underset{i \in N}{\arg\max}(T_i) - \underset{i \in N}{\arg\min}(T_i) \right\|$. Testing data clusters of various curvatures led to the selection of

234   approximate curvature threshold to 0.15. The segments were assigned as curved using

$$Curved = \begin{cases} true & if \ (T \in [-R,+R] \vee T \in [+R,-R]) \wedge Curvature = 1 \\ false & Otherwise \end{cases} \tag{3}$$

235   where $Curvature = \begin{cases} 1 & if \ \eta \geq 0.15 \\ 0 & Otherwise \end{cases}$ and R is the quantitative measure of the change in y (**Fig. 3**).
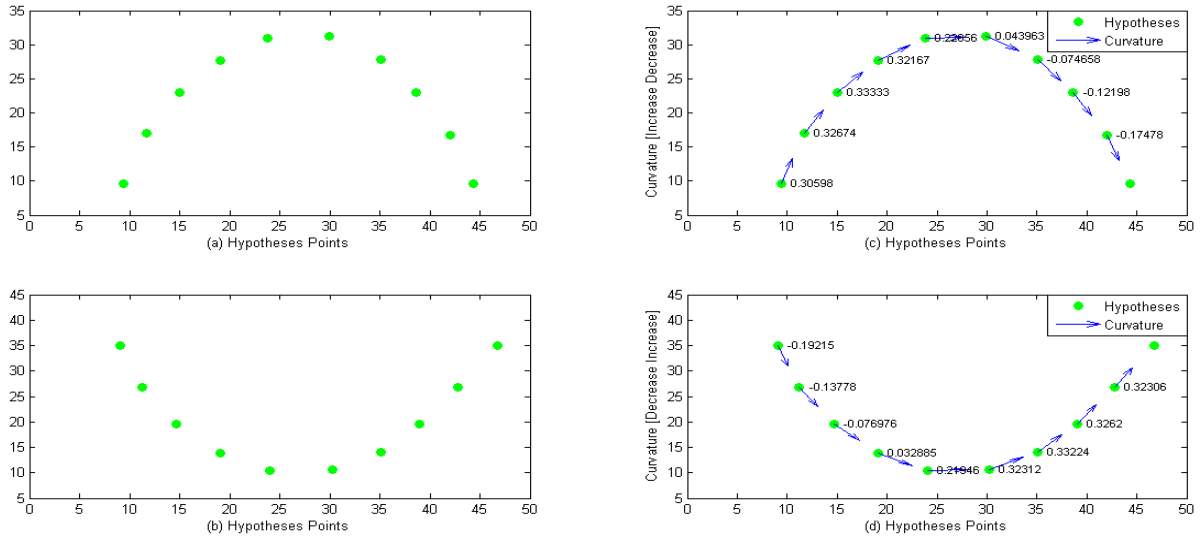
236
237   **Figure 3:** Clusters with curvature of hyperbolic features

238

239   The curvature detection algorithm was tested on various groups of hypotheses which

240   demonstrated robust performance in terms of separating curved segments from linear ones (four

241   of the groups given in **Fig (4)**)
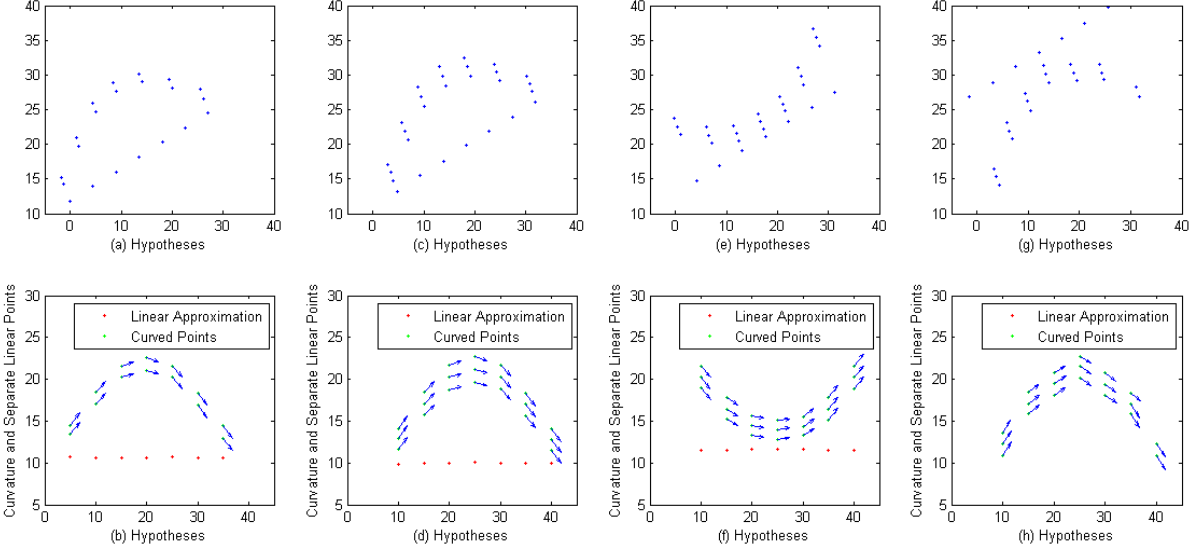
Figure 4: Separation of hypotheses into cluster of linear clusters and curved clusters

For clusters with circular features, a three point based was used as given in **Fig A1 (Appendix).**

2.3. Segment-Manhole Connection using Expectation Maximization Algorithm

The connection of segments recognized from previous section with manholes was categorized as a classification problem where each segment was to be classified as a segment connected to two manholes. The EM algorithm was proposed for segment-manhole connection establishment which has been used in wide range of machine learning and data mining scenarios such as classification (Aldebaro Klautau, 2003, Sander et al., 1998), Image Processing (Huanhuan Chen, 2010) and unsupervised clustering (Bailey and Elkan, Buntine, 2002, Salojarvi et al., 2005). The inputs from sensor readings, estimated asset directions from manholes, statutory records and segments recognized from previous step were used to identify suitable connections between segments and manholes. Combining the inputs created a data fusion scenario where hidden information from different sources was integrated and the probability of segment classification was updated until the algorithm converged (or the number of iterations reaches). Therefore, based on the probability distribution drawn from EM algorithm, the local maximum of segment-manhole connection

14

258 provided an optimal solution that was then compared with prior information from statutory

259 records. Using EM, each segment can be assigned to a pair of manhole connections given the

260 above parameters. EM consists of two steps that are (1) Expectation and (2) Maximization.

261 Expectation step calculates the probability distribution of a target outcome given a set of

262 parameters while the maximization step updates the distribution given the parameters which are

263 updated from expectation step at each iteration. The EM algorithm is explained as follows:

264 1) For a given hypothesis point h, get nearest segment S_j from statutory record. Get manhole
265   locations and the depths. Initialise the prior for each manhole connection with P_i as uniform
266   distribution. Initialise covariance matrix
267     $\Lambda = diag(\Delta d, \|\theta_j - \theta_k\|, \|\theta_k - \theta_{hk}\|, \|\theta_j - \theta_{hk}\|, \|Z_h - Z_k\|)$

268   In the above equation, $\Delta d = \sqrt{(x_{Sj} - x_k)^2 + (y_{Sj} - y_k)^2}$, θ_j = direction of S_j in the map, $\theta_k =$
269   direction of segment of same type as of S_j recorded from manhole k's cover, θ_{hk} = direction
270   from manhole k to h, z_h = depth of h recorded by the sensor, z_k = depth of the segment
271   of same type as of S_j recorded from manhole k's cover, and $\Theta = (K, X, Y, Z, x_h, y_h, z_h, S_j)^T$ for
272   K manholes with $[X, Y, Z]$ as position and depth information.

273 2) E-step: Likelihood for each manhole k connection

274 $$L\langle C_k | \Lambda, \Theta \rangle = \frac{1}{(2\pi |\Lambda|)^{3/2}} \exp\left\{-\frac{1}{2}\left(C(h,k)^T \Lambda^{-1} C(h,k)\right)\right\}$$

275   Where $C(h,k) = \left[\Delta d, \|\theta_j - \theta_k\|, \|\theta_k - \theta_{hk}\|, \|\theta_j - \theta_{hk}\|, \|z_h - z_k\|\right]$. The Maximum Likelihood Estimate
276   (MLE) is then given by

277 $$\gamma_k = \frac{L\langle C_k | \Lambda, \Theta \rangle}{\sum_{k=1}^{K} L\langle C_k | \Lambda, \Theta \rangle}$$

278 3) M-Step: Maximisation of likelihood
279

280 $$\Lambda = diag\left(\begin{array}{cc} \sqrt{\dfrac{\left(\sum_k \prod_{k=1}^{K}(\Delta d. L\langle C_k | \Lambda, \Theta \rangle)\right)^2}{\sum_{k=1}^{k} L\langle C_k | \Lambda, \Theta \rangle}}, & \sqrt{\dfrac{\left(\sum_k \prod_{k=1}^{K}(\|\theta_j - \theta_k\|. L\langle C_k | \Lambda, \Theta \rangle)\right)^2}{\sum_{k=1}^{k} L\langle C_k | \Lambda, \Theta \rangle}}, \\[30pt] \sqrt{\dfrac{\left(\sum_k \prod_{k=1}^{K}(\|\theta_k - \theta_{hk}\|. L\langle C_k | \Lambda, \Theta \rangle)\right)^2}{\sum_{k=1}^{k} L\langle C_k | \Lambda, \Theta \rangle}}, & \sqrt{\dfrac{\left(\sum_k \prod_{k=1}^{K}(\|\theta_j - \theta_{hk}\|. L\langle C_k | \Lambda, \Theta \rangle)\right)^2}{\sum_{k=1}^{k} L\langle C_k | \Lambda, \Theta \rangle}}, \|z_h - z_k\| \end{array} \right)$$

15

281

282   4) Repeat 2-3 until converged (log likelihood converges to 10^{-4})

283

284   5) Maximum A Posteriori (MAP) for connection of k with S_j

285         $C_{h\to k} = \arg\max_{k\in K}(\gamma_k)$

286 The MLE γ_k in step 2 provides best match between a manhole k and a segment S_j. The algorithm

287 is repeated for connection from both ends of S_j to a pair of best matching manholes. At the test

288 site, it is possible for sensors to contain noise (containing spatial location and direction error) as

289 part of the hypotheses. A point h was considered noise when it's likelihood of manhole

290 connections ($k_{\to}$ and $k_{\leftarrow}$) fell below a threshold

291 $\alpha = \dfrac{\sum_{h\in O} P(L|h,k_{\to},k_{\leftarrow})}{\frac{3}{2}O}$ where $P\big(L|h,k_{\to},k_{\leftarrow}\big) \approx N(L,h,\sigma_o)$

292 In the above equation, $\sigma_o = 0.4$ is the variance (meters) calculated from (x,y,z) of O. The noise

293 from above set of hypotheses points was identified using

$$h = \begin{cases} noise & if \ P\big(L|h,k_{\to},k_{\leftarrow}\big) < \alpha \\ h & otherwise \end{cases} \tag{4}$$

294 The clusters of segments for manhole connections from (4) consisted of the hypotheses points and

295 the connected manhole positions which are:

296 $C_j = \big[k_{\to} \quad \{h\in O : h\to S_j\} \quad k_{\leftarrow}\big]$

297   **3. Results**

298   Dataset for model validation included five sites of simulated data (3 sets) as well as real data (2

299 sets). The model was developed and validated using MATLAB 2011(b) where the most probable

300 maps were generated given 3 different sources of information: (i) sensor readings, (ii) manhole

301 surveys, and (iii) the statutory records. In order to validate the model, hypotheses extraction,

302 accurate segment identification (with noise removal) and segment-manhole connection were

16

303 tested. Several variations in the parameters and data quality were taken into account in order to

304 verify the robustness of model. The hypotheses were extracted from raw sensor images using

305 image segmentation techniques. A segment recognition algorithm was developed to identify linear

306 and curved pipe/cable segments based on the direction of survey taken by MTU machine, and

307 directions estimated from hypotheses when they were combined together. Total of 4 initial use

308 cases of hypotheses (**Fig. 5**) tested in order to evaluate the performance of segment recognition

309 algorithm where the number of segments in each use case were known.
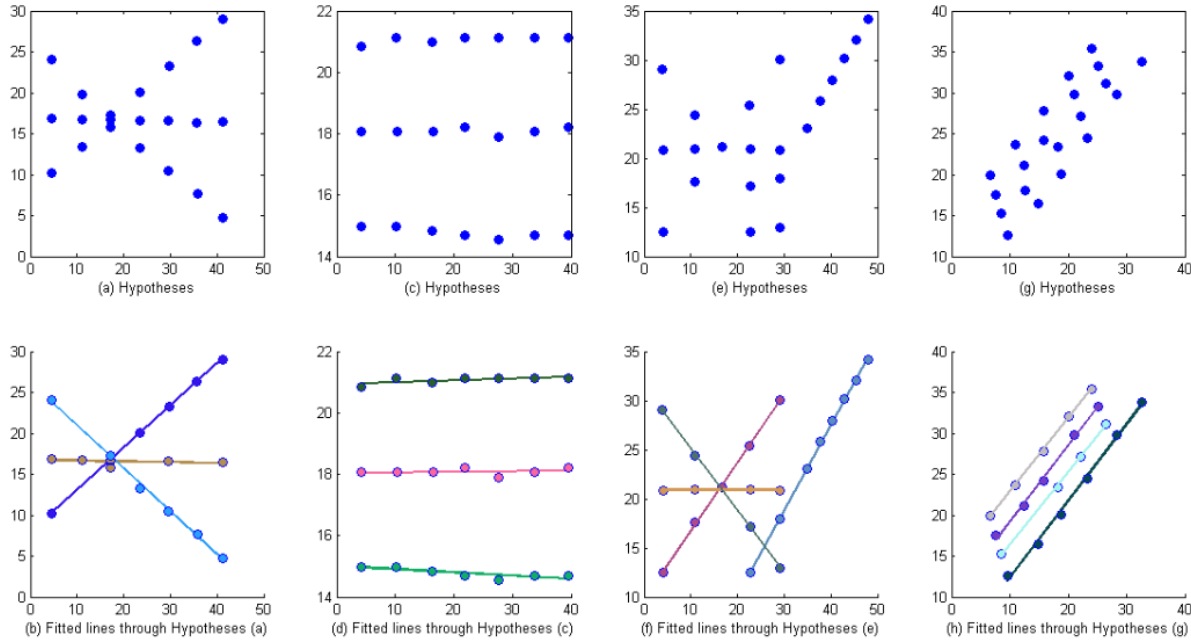
Figure 5: Segment recognition algorithm fitting linear segments in groups of hypotheses

310 The simulated datasets (S1, S2, and S3) consisted of locations of manholes as well as pipes from

311 ArcGIS (Esri, 2015) and simulated hypotheses (simulated readings from MTU sensors). The real

312 sites (R1 and R2) consisted of measurements from MTU sensor readings and statutory records.

313 For noise removal in simulated sets, white noise was added with the following: (1) Spatial noise

314 of the locations of manholes and pipes was up to 2×2 meters, (2) Noise in hypotheses locations

315 was up to 0.4 meters and (3) Noise in pipe directions from manholes was up to 8 degrees.

316 **Table 1:** Settings for simulated and real sites (S# = simulated site #, R# = real site #)

| Data	# Manholes	# Pipes & Cables	# Sensor Readings
S1	10	13	28
S2	12	16	30
S3	26	46	87
R1	19	8	18
R2	8	12	25

317

318 **3.1. Bayesian Mapping Model for Simulated Data:**

319 The model was initially tested on 3 simulated datasets (S1, S2 and S3) with varying numbers of

320 manholes, sensor data points, and segments from statutory records (**Table 1**). For simulated

321 datasets, the hypotheses were manually generated and the initial maps were obtained from ArcGIS

322 (Esri, 2015). Initial simulated datasets with initial maps are depicted in **Fig 6**.
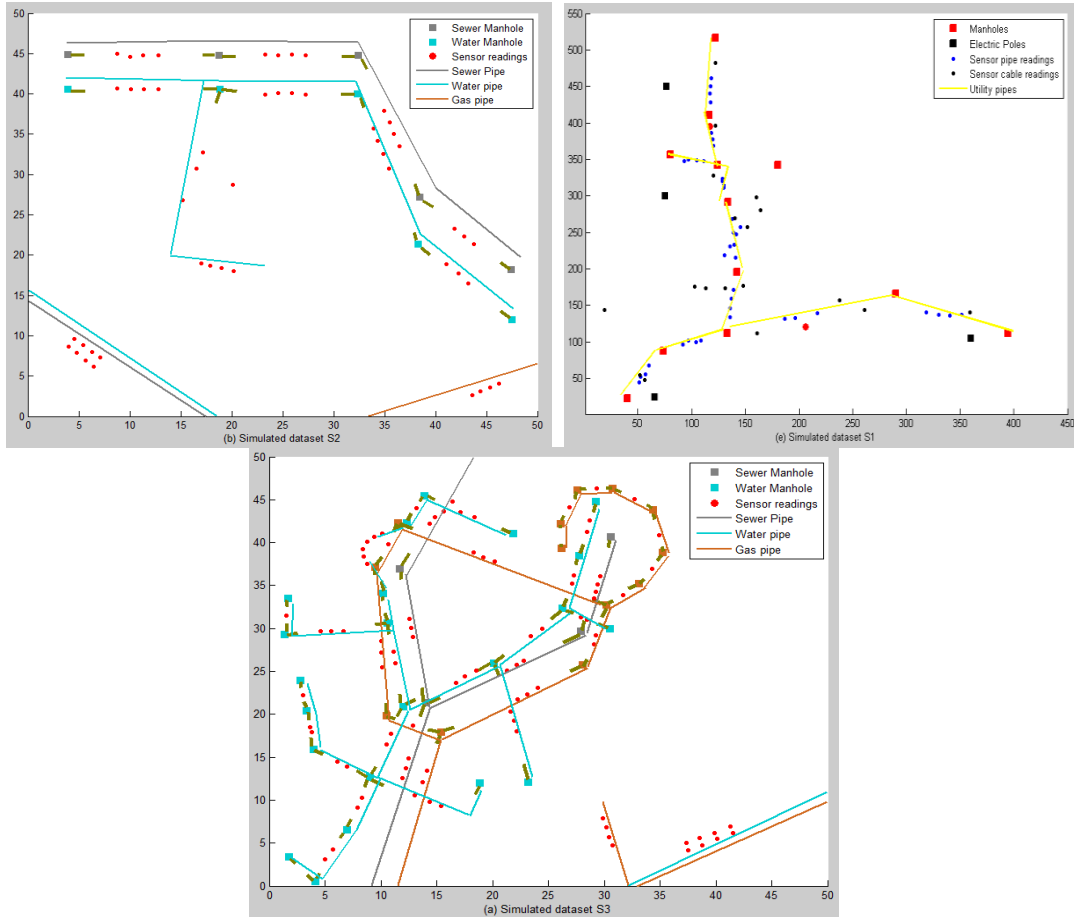

325 **Figure 6:** Simulated datasets with hypotheses points and initial maps **(a)** for S1, **(b)** for S2, and
326 **(c)** for S3

327   **(a) Segment Recognition:** Segment recognition algorithm showed robust performance for all

328   datasets and the noise was removed from groups of hypotheses in S1 and S2. The absence of

329   statutory records did not affect the capability of the model to draw accurate segments due to the

330   segment recognition algorithm. In addition, the segments were drawn in 3D due to the inclusion

331   of depth information from manholes and sensor readings. However the uncertainty increased in

332   the absence of statutory records as they provided useful information on the number of buried assets

333   and their directions. The segments drawn by automated segment recognition algorithm for

334   simulated datasets are given in **Fig 7.**

335
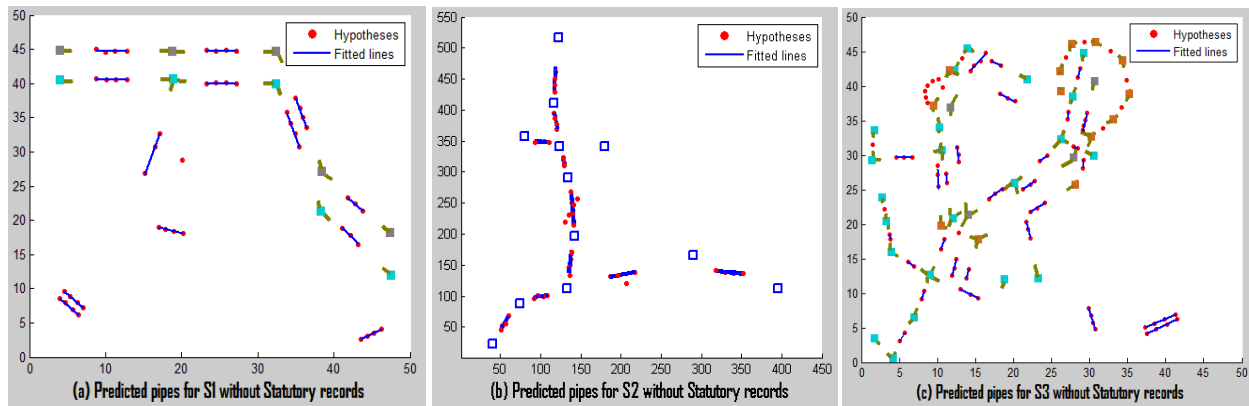
336   **Figure 7:** Segments recognized for S1 **(a)**, S2 **(b)** and S3 **(c)**

337   **(b) Segment-Manhole Connection:** The segment recognition algorithm provided advantages

338   over (Chen and Cohn, 2011) in terms of basis for accurate connection with manholes and

339   construction of 3D maps. The segments were used by EM algorithm for segment-manhole
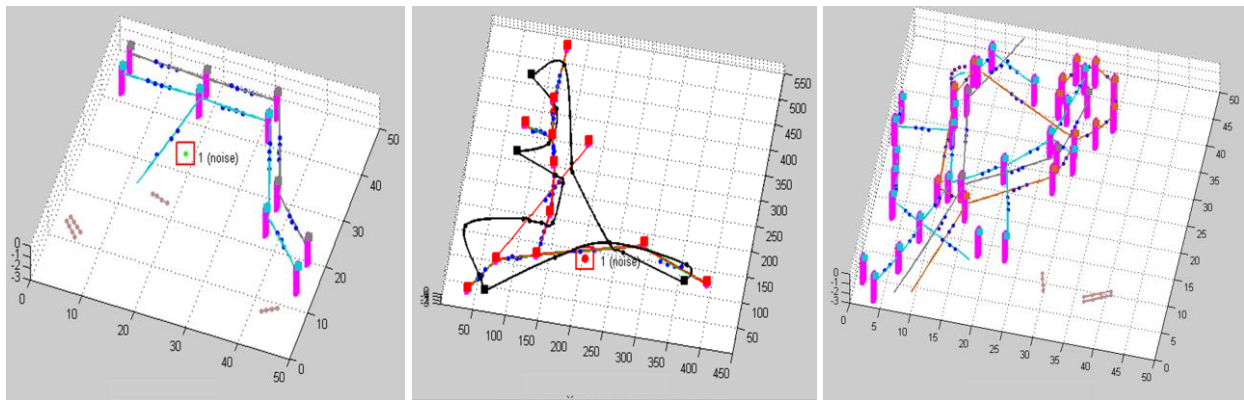

Figure 8: 3D map construction for segment-manhole connection (S1**(a)**, S2**(b)** and S3**(c)**)

19

340 connection establishment. Using the parameters from different sources (Materials and Methods),

341 it was observed that the algorithm successfully created connections for segments that led to the

342 reconstruction of 3D map. Among the segments generated from step (a) a few segments had single

343 end connecting to a manhole. Such situations were also tested for the validation of EM algorithm.

344 The maps of simulated datasets were generated which are depicted in **Fig 8**.

345 **3.2. Bayesian Mapping Model for Real Data**

346 Real sites (R1 and R2) with sensor readings and manhole surveys were tested for model

347 validation. **Table 1** provides information on the inputs to model for simulated sites as well as real

348 sites. There were 18 sensor readings taken at R1 which included readings for linear pipes and an

349 electricity cable. Raw images from sensors were manipulated to extract hypotheses which were

350 used for automatic segment recognition. There were no connection errors for R1 using EM based

351 segment-manhole connection algorithm for linear segments.
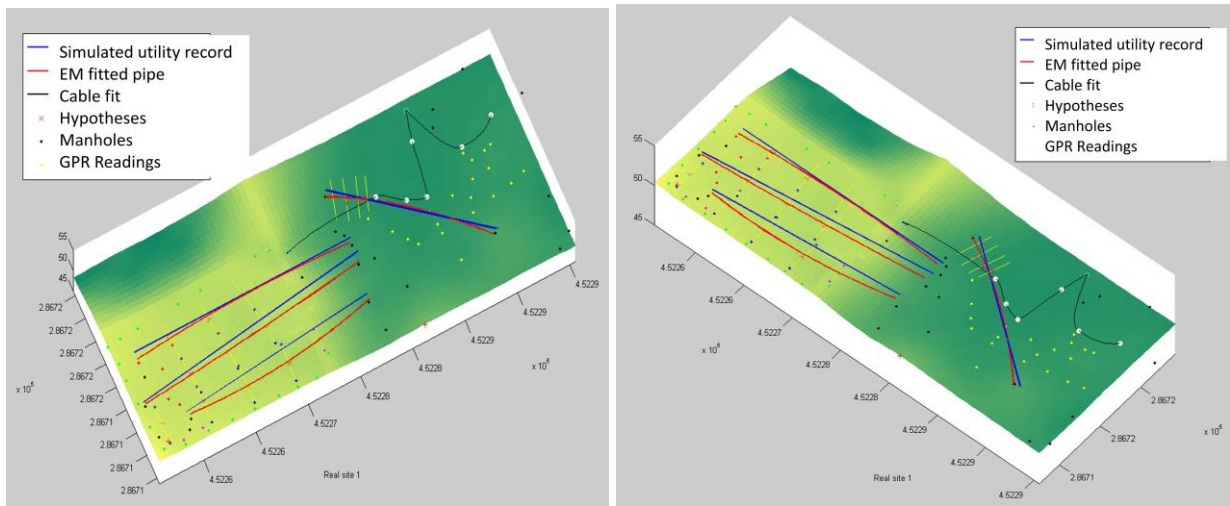
Figure 9: Real Site 1. Blue lines represent simulated utility record based on manhole and sensor readings. Red lines represent segment-manhole connections established by the model.

352 Real site R2 consisted of straight linear pipes with an electric cable for which the sensor readings

353 were taken (**Fig 10a**). There was no noise detected in the sensor readings and segments were

354 successfully connected with manholes. PMF sensor usually detects objects with electrical current

355　such as electric cables. At R2, The PMF sensor detected a water pipe (rectangle in **Fig 10b**) in the

356　survey. This was due to the run of an electric cable in a close proximity of that water pipe which

357　emits the electrical current.

358

359
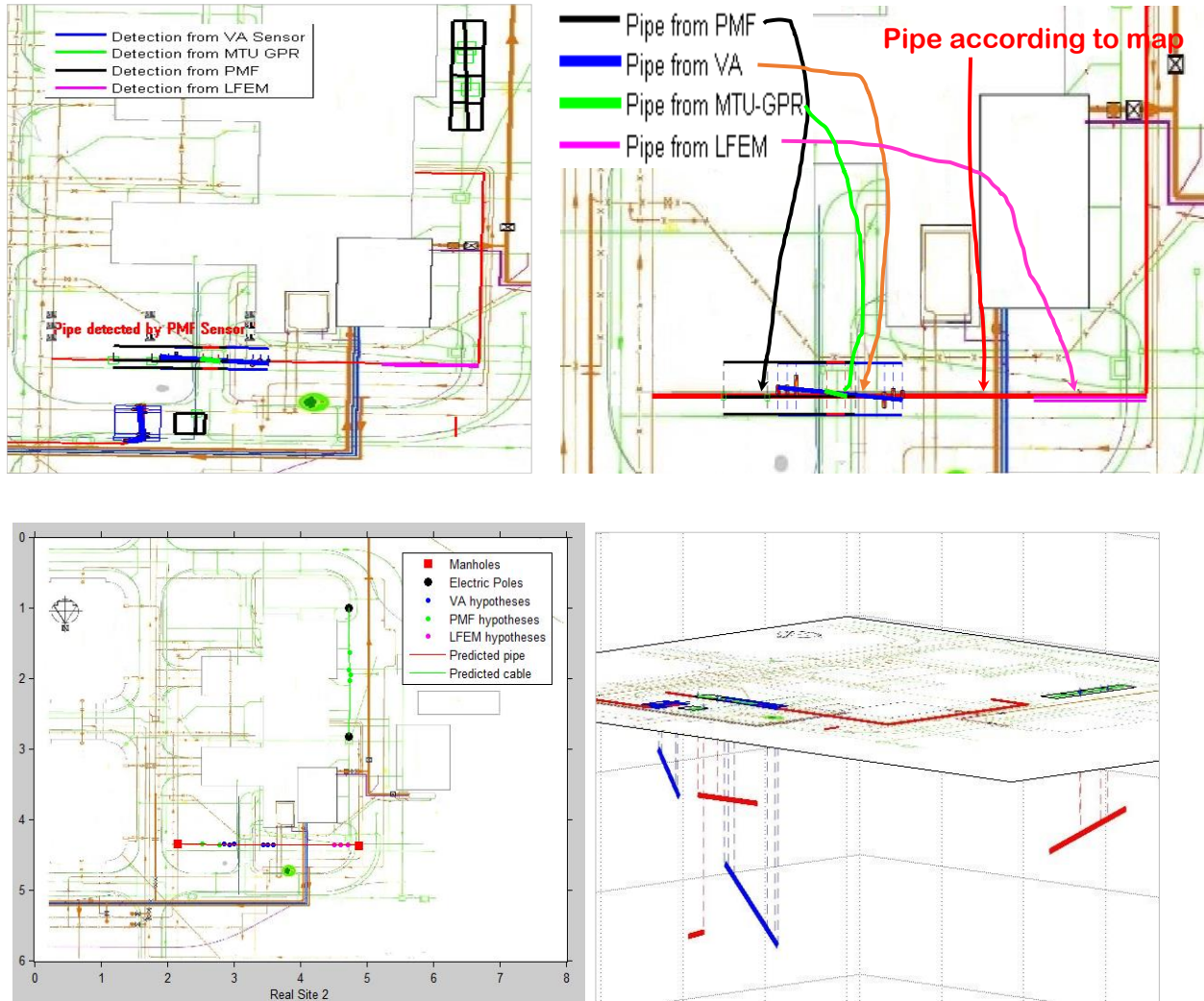
360　**Figure 10:** Real Survey Site 2 **(a)** pipe detected by MTU sensors, **(b)** pipe detected by PMF sensor
361　due to electric cable buried closely, **(c)** data from different sensors integrated together for segment-
362　manhole connection, **(d)** 3D view of segments detected by Bayesian Mapping model

3.3. Connection and Spatial Errors in Mapping Models

364　For manhole connection establishment process, the proposed algorithm was also compared with

365　Joint Compatibility Branch & Bound (JCBB) which calculates the Mahalanobis distance between

366　the observations and the predictions and accepts a connection if the Mahalanobis distance is

21

367   smaller than a validation gate (Yangming Li, 2013). The segment-manhole connection problem

368   was similar to the spatial data association problem in robotics (Bailey, 2006). Using the data fusion
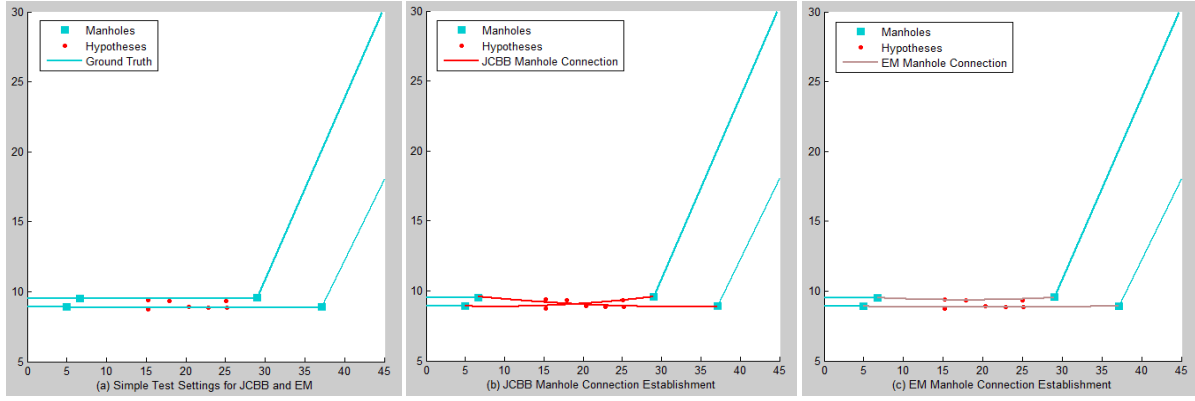
Figure 11: Manholes closed to each other potentially suitable for connection with one segment

369   technique, noise removal and the iterative refinement of posterior probability using EM algorithm,

370   the analysis showed better results compared to JCBB for both simulated sets as well as real sites.

371   JCBB suffered in connecting segments in simple scenarios with manholes when more than one

372   segments were located in close proximity in the map as shown in **Fig 11**.

373   In tested datasets, JCBB also suffered from connection errors at the manhole locations where

374   directional error exceeded five degrees. There were total five manhole connection errors recorded
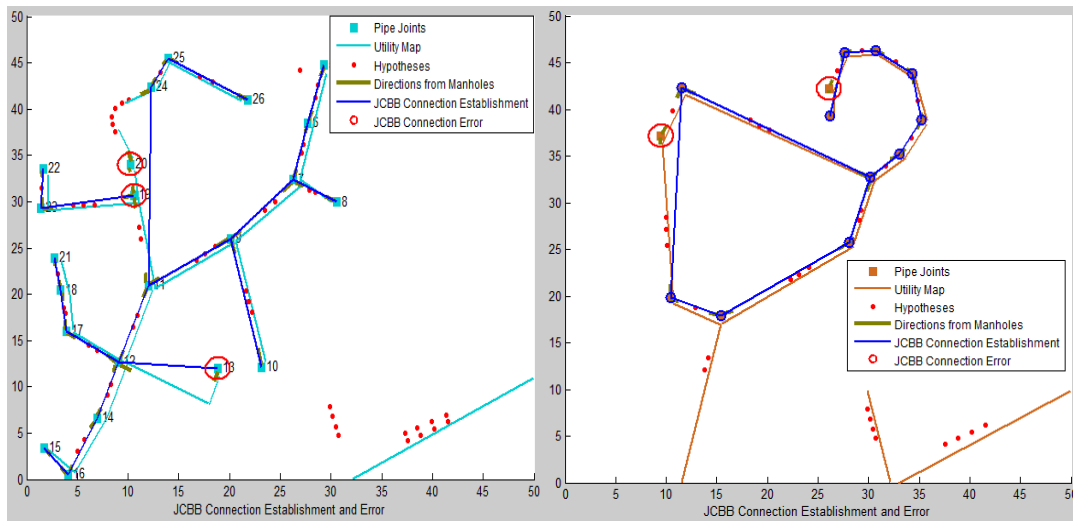
Figure 12: (a) JCBB Connection Error in water pipes **(b)** JCBB Connection Error in gas pipes

22

375 when JCBB was run on S3 as shown in **Fig 12** where **Fig 12a** shows the connection errors for

376 water pipes and **Fig 12b** shows the connection errors for gas pipes.

377   The analysis with segment recognition algorithm significantly reduced the spatial and directional

378 errors when modeling for mapping the underground utilities. The mean directional error using EM

379 algorithm was smaller compared to JCBB. The connection errors are given in **Table 2** which are

380 denoted as E(JCBB) for JCBB and E(EM) for EM algorithm. Table 2 below shows the results of

381 proposed EM algorithm with JCBB.

382 **Table 2:** Directional $(D(\Theta))$, Spatial $(S(x,y))$ and connection errors compared to JCBB. **EM\L** =
383 without segment recognition algorithm, **EM\U** = without utility records

Error	EM\L	EM\U	EM	JCBB	U	E(JCBB)	E(EM)
$S(x,y)_{S1}$	1.3514	1.7923	0.4132	3.1548	2.8	1	0
$D(\Theta)_{S1}$	3.3371	4.4451	1.2526	6.1112	4		
$S(x,y)_{S2}$	2.1334	3.5691	1.005	7.4324	2.5	0	0
$D(\Theta)_{S2}$	4.3351	6.4321	1.1129	10.9813	5		
$S(x,y)_{S3}$	3.1058	6.9471	2.0591	12.6976	6.5	5	0
$D(\Theta)_{S3}$	4.6131	8.8195	2.0031	16.9115	6		
$S(x,y)_{R1}$	4.3291	8.7795	1.6531	11.1125	5	1	0
$D(\Theta)_{R1}$	3.9973	6.9553	1.8533	8.8134	4		
$S(x,y)_{R2}$	2.4415	3.1101	1.0051	4.9518	3	0	0
$D(\Theta)_{R2}$	2.3705	3.1781	0.8109	4.1136	3		

384

385   A noticeable impact on the voxel prediction for each pipe was observed when this error rate was

386 increased from 8 to 14 degrees. The increased directional error in S1 and S3 had significant effect

387 with inaccurate segment-manhole connections in the maps. In addition, the spatial error resulted

388 in variations in voxel classification as shown in **Fig 13a** and **Fig 13b**. The black circles show

389 directional errors and the red circles show spatial error.

390
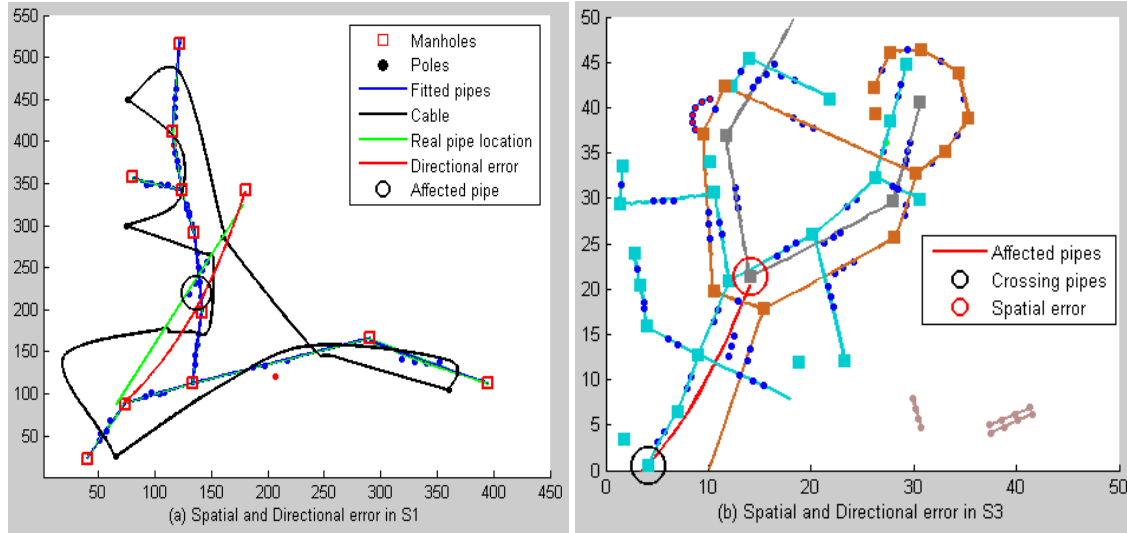
Figure 13: Impact of spatial and directional error on survey sites S2 and S3

4. Discussion and Conclusions

The traditional approaches for map generation given in the literature require information from statutory records and provide solutions for mostly 2D map generation. Additionally, the information from multiple sensors require extensive processing and statutory records should also be integrated that provide prior knowledge about buried utilities. Recent analyses/studies on the abilities to locate underground utilities stressed the needs of a system capable of generating real time maps with the help of heterogeneous information. However, current techniques are limited to either generating 2D maps of linear segments or unable to detect different types of utilities. This paper addresses these problems by proposing a Bayesian mapping model by implementing various machine learning techniques for real time 3D map (re)construction. The segment recognition algorithm is robust in identifying groups of hypotheses forming linear/curved segments that are helpful in establishing connections with manholes. In order to improve utility classification and refined map generation, noise removal facilities were embedded in the system that improved performance in distinguishing between hypotheses and noise. The Bayesian Mapping model is aimed to overcome critical issues related to efficient and real time location of buried assets that

24

406 could provide valuable underground information and be cost effective. Further online analysis will

407 also validate model performance at higher levels and its ability to generate refined maps at real
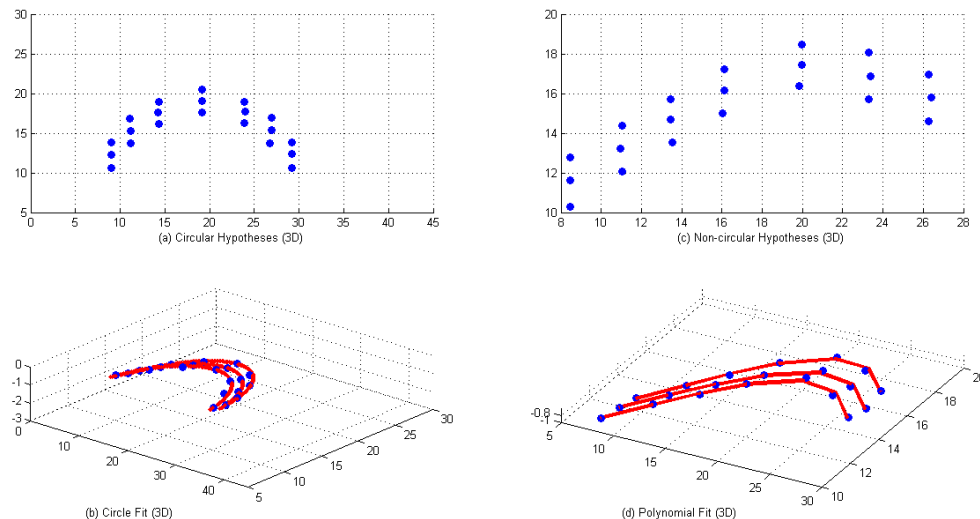
408 time.

Acknowledgements

412

413 **Appendix**

414 For curved segments, the three point circle fitting was applied as illustrated in Fig. A-1

415
416 **Figure A1: (a)** Circular hypotheses, **(b)** Circle fit for circular points, **(c)** Non-circular Hypotheses,
417 **(d)** Polynomial fit for non-circular points
418

419 **A. Most Probable segment estimation**

420 In order to refine the output of linear segment fitting without significant impact on model

421 performance, the least squares fitting was performed on the above sets of hypotheses and manhole

422 connections. Even though the data from multiple sources was integrated for Bayesian mapping,

423 the sensor readings may contain small non-negligible noise which may also affect the performance

424 of least squares linear fitting algorithm. **Figure A-2** shows a simple scenario where the point

425 containing noise in its location causes an increased angular difference with the direction from

426 connecting manhole. Also the segment approximation fails to find the most probable linear

427 segment without crossing any other segment in close proximity.

428
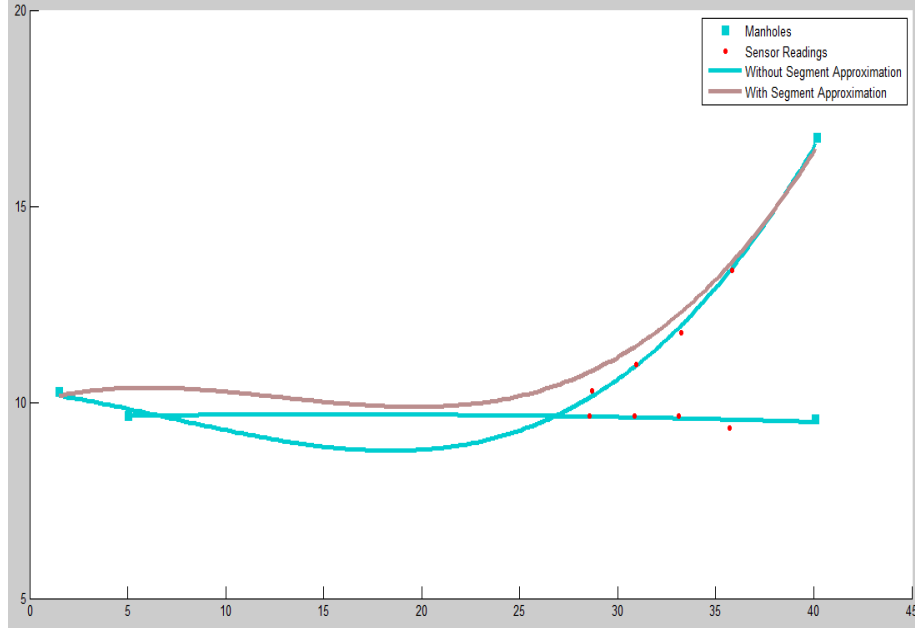
Figure A-2: Pipe segment fitted from hypotheses crossing another closed segment in the map

429

430   It is critical to develop an approach for above mentioned scenarios since it is common that the

431   street surveys contain more than one pipes buried in close proximity and the chances of having the

432   overlaps between those segments are higher. In order to overcome such challenge, a probability

433   distribution over the space of three possible lines L_i $(i = 1,2,3)$ was created which was described by

434   discrete random variables. L_i is the ith line of quantized voxels given the specific attributes which

435   are the locations of the points of clusters defined in previous section. We need to find

436
$$\arg \max_{l} \prod_{l=1}^{g_l} \left(P\left(y_l \middle| L_1, L_2, L_3\right) \right) \qquad\qquad (A1)$$

437   and

438
$$\arg \max_{l} \prod_{l=1}^{g_l} \left(P\left(z_l \middle| L_1, L_2, L_3\right) \right) \qquad\qquad (A2)$$

439   Where $l \in g_l$ is the voxel chosen from a set g_l of voxels fitted for the quantized line. The probability

440   distribution for three observed possible 3D positions of each line L_i were created and combined

27

441   given x, y and z information. For each voxel of a line L_i, two equal spaced perpendicular line

442   segments g_p were drawn along the y-axis for (A1) and z-axis for (A2). This set of linear segments

443   is denoted by G. For the calculation of the most probable voxel at each quantised position of each

444   line L_i, we use two terminologies. For (A1), the pixel p_i^l of L_i is the (x_l, y_l) and for (A2), the pixel

445   p_i^l of L_i is the (y_l, z_l). Therefore (A1) is used for determining the most probable y^{th} position of

446   the voxel and (A2) is used to determine the zth position of the voxel in the map. For all k voxels

447   defined by g_l, the joint probability distributions are given by

448    $$P(y_l|L_1, L_2, L_3) = \prod_{i=1}^{L} P(y_l|L_i) \; for \; L = 3$$

449    $$P(z_l|L_1, L_2, L_3) = \prod_{i=1}^{L} P(z_l|L_i) \; for \; L = 3$$

450   Where $$P(y_l|L_i) = \sum_{l=1}^{g_l} \sum_{y \in G} \frac{1}{\|\nabla y\|} \exp\left\{ -\frac{\|y - p_i^l\|^2}{\|g_p\|^2} \right\}$$ (A3)

451   p_i^l is the lth pixel (x_l, y_l) of L_i

452    $$P(z_l|L_i) = \sum_{l=1}^{g_l} \sum_{z \in G} \frac{1}{\|\nabla z\|} \exp\left\{ -\frac{\|z - p_i^l\|^2}{\|g_p\|^2} \right\}$$ (A4)

453   p_i^l is the lth pixel (y_l, z_l) of L_i

454   Calculating $P(y_l|L_i)$ and $P(z_l|L_i)$ provide the probabilities $P(y_l|L_1, L_2, L_3)$ and $P(z_l|L_1, L_2, L_3)$

455   respectively which are used to calculate the most probable voxel position for the map using

456
$$P(L_y|L_1,L_2,L_3) = \arg\max_l \prod_{l=1}^{g_l}\left(P(y_l|L_1,L_2,L_3)\right)$$

(A5)

457 and

458
$$P(L_z|L_1,L_2,L_3) = \arg\max_l \prod_{l=1}^{g_l}\left(P(z_l|L_1,L_2,L_3)\right)$$

(A6)

459 An example of this is shown in **Figure A-3** where three lines (L_1,L_2,L_3) are drawn from three

460 techniques and the quantized lines on y-axis and z-axis are drawn to calculate (A5) and (A6).

461
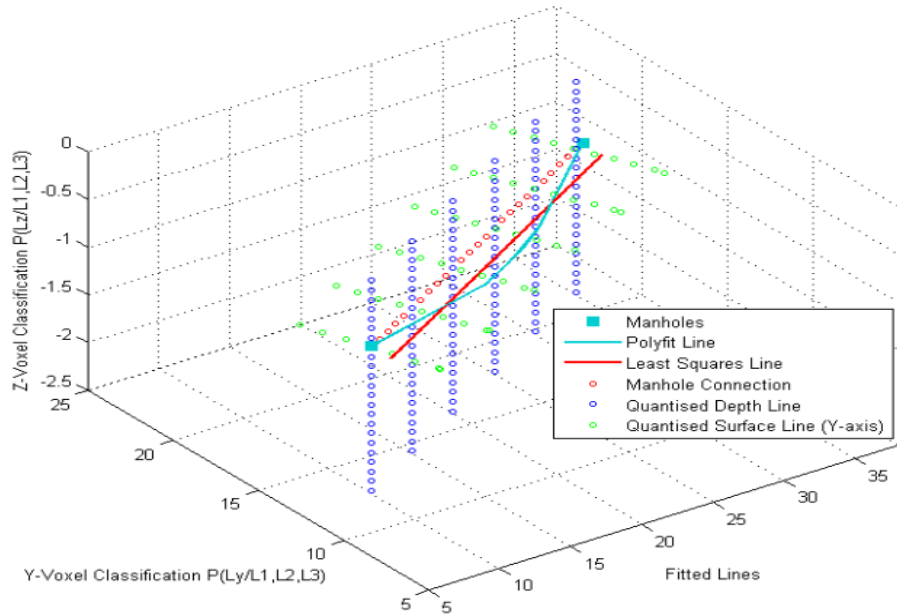
462   **Figure A-3**: Perpendicular quantised linear points created from the point and the linear segment

463 At each step in the map, the voxel with the highest probability given three lines as input was taken

464 as the most probable voxel. The set of all voxels at the end created a line in 3D space which was

465 considered as the most probable linear segment approximation. This algorithm was applied for the

466 approximation of each linear segment in the performance evaluation of the algorithm.

467

468

29

B. Cardinal Spline Fitting

In addition to the linear and curved pipe segments which may include sewer pipes, gas pipes and water pipes, the street maps also included electric cables buried underground which need to be approximated. The difference of the cables from the other pipe segments is their non-linear behavior. The cable can be buried in an arbitrary order without satisfying linearity condition for which line fitting or curve approximation may represent inappropriate solutions. The Passive Magnetic Field (PMF) sensors report the sequence of the survey locations when performing street survey. For fitting cables and non-linear segment, Cardinal Spline algorithm was implemented which was capable for fitting the lines to points in both 2D and 3D. The detailed explanation and the functionality of the Cardinal Spline algorithm is provided elsewhere (Ali Khan and Sarfraz, 2011).

C. Probabilistic Voxel Classification

Errors in the linearity of segment lines vary depending on the noise added in sensor and manhole readings. In addition, the presence of multiple closely-buried assets is challenging when recognizing the number of segments. In order to address these challenges, linear approximation algorithm was useful as it integrated information from multiple sources and provided most probable estimation of voxel classification. Bayesian probabilistic voxel classification showed improved results in terms of fitted lines when compared to least squares fitting.
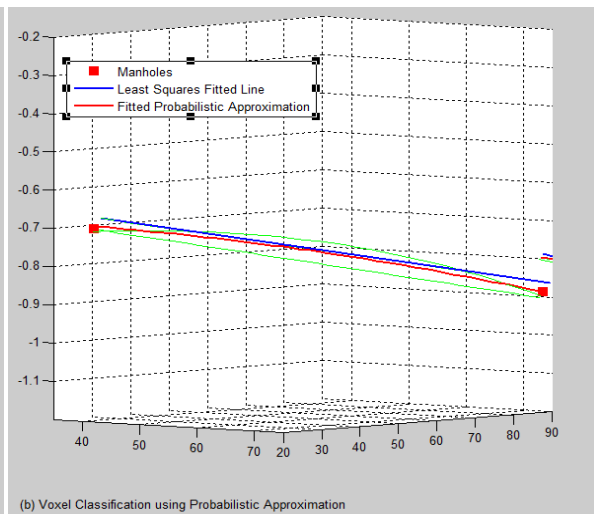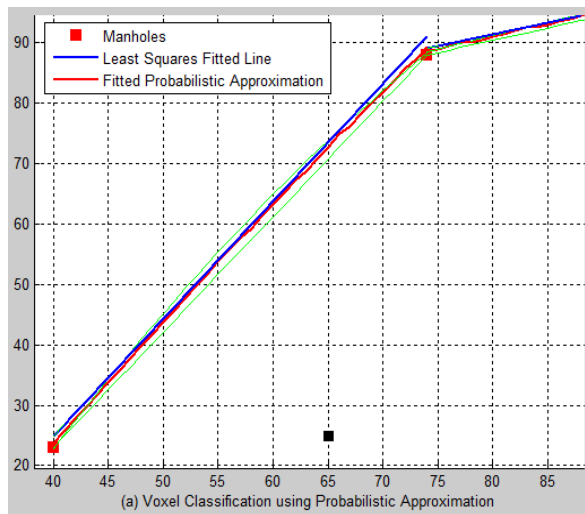
487
488 **Figure C-1**: Lines regressed for two pipes with and without using probabilistic approximation
489 algorithm

References

abhir Bhalerao , R. W. A Fourier Approach to 3D Local Feature Estimation from Volume Data (2001). British Machine Vision Conference, 2001. 461-470.

Aldebaro Klautau, N. J. A. A. O. 2003. Discriminative Gaussian Mixture Models: A Comparison with Kernel Classifiers. In: MISHRA, T. F. A. N. (ed.) Proceedings of the 20th International Conference on Machine Learning (ICML-03).

Ali Khan, M. & Sarfraz, M. 2011. Motion Tweening for Skeletal Animation by Cardinal Spline. In: Abd Manaf, A., Sahibuddin, S., Ahmad, R., Mohd Daud, S. & El-Qawasmeh, E. (eds.) Informatics Engineering and Information Science. Springer Berlin Heidelberg.

Ashdown 2000. Mains Location Equipment – A State of the Art Review and Future Research Needs. In: UKWIR (ed.).

Bailey, T. D.-W., H. 2006. Simultaneous localization and mapping (SLAM): part II. IEEE Magazine on Robotics & Automation, 108 - 117.

Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers.

Bar-Shalom, Y. 1987. Tracking and data association, Academic Press Professional, Inc.

Buntine, W. 2002. Variational Extensions to EM and Multinomial PCA. In: Elomaa, T., Mannila, H. & Toivonen, H. (eds.) Machine Learning: ECML 2002. Springer Berlin Heidelberg.

Chen, H. & Cohn, A. G. 2011. Buried utility pipeline mapping based on multiple spatial data sources: a bayesian data fusion approach. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three. Barcelona, Catalonia, Spain: AAAI Press.

Delicado, P. & Smrekar, M. Mixture of nonlinear models: a Bayesian fit for Principal Curves.

Do, C. B. & Batzoglou, S. 2008. What is the expectation maximization algorithm? Nat Biotech, 26, 897-899.

Dong-Min, W. & Dong-Chul, P. 3D Line Segment Detection Based on Disparity Data of Area-Based Stereo. Intelligent Systems, 2009. GCIS '09. WRI Global Congress on, 19-21 May 2009 2009. 219-223.

Esri. 2015. Put Your Maps to Work with ArcGIS, the Mapping Platform for Your Organization [Online]. Esri. Available: https://www.arcgis.com [Accessed 10/04 2015].

Ester, M., Kriegel, H.-P., J╠Â╤║Rg, S. & Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. 226-231.

Friedman, J. & Popescu, B. 2004. Gradient Directed Regularization for Linear Regression and Classification.

Huanhuan Chen, A. G. C. 2010. Probabilistic Robust Hyperbola Mixture Model for Interpreting Ground Penetrating Radar Data. The 2010 International Joint Conference on Neural Networks (IJCNN). Barcelona: IEEE.

M Burtwell, E. A. 2004. Locating Underground Plant and Equipment - Proposals for a Research Programme.

Mcmahon, W., Burtwell, M.H., Evans, M. 2005. Minimising Street Works Disruption: The Real Costs of Street Works to the Utility Industry and Society. In: UKWIR (ed.).

Neira, J., And Tardos, J.D. 2001. Data association in stochastic mapping using the joint compatibility test. IEEE Transactions on Robotics ad Automation, 890-897.

Royal, A. C. D., Atkins, P. R., Brennan, M. J., Chapman, D. N., Chen, H., Cohn, A. G., Foo, K. Y., Goddard, K. F., Hayes, R., Hao, T., Lewin, P. L., Metje, N., Muggleton, J. M., Naji, A., Orlando, G., Pennock, S. R., Redfern, M. A., Saul, A. J., Swingler, S. G., Wang, P. & Rogers, C. D. F. 2011. Site Assessment of Multiple-Sensor Approaches for Buried Utility Detection. International Journal of Geophysics, 2011, 19.

Royal Acd, R. C., Atkins Pr, Brennan Mj, Chapman Dn, Chen H, Cohn Ag, Curioni G, Foo Ky, Goddard K, Hao T, Lewin Pl, Metje N, Muggleton Jm, Naji A, Pennock Sr, Redfern Ma, Saul Aj, Swingler Sg And Wang P 2010. Mapping the Underworld: Location, Mapping and Positioning without Excavation. In: CONTROL, C. (ed.). Middle East, Singapore.

Salojarvi, J., Puolamaki, K. & Kaski, S. 2005. Expectation maximization algorithms for conditional likelihoods. Proceedings of the 22nd international conference on Machine learning. Bonn, Germany: ACM.

Sander, J., Ester, M., Kriegel, H.-P. & Xu, X. 1998. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Min. Knowl. Discov., 2, 169-194.

Sanquer, M., Chatelain, F., El-Guedri, M. & Martin, N. A reversible jump MCMC algorithm for Bayesian curve fitting by using smooth transition regression models. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 22-27 May 2011 2011. 3960-3963.

Ting-Wang, C. & Qing, W. 3D line segment detection algorithm for large-scale scenes. Audio Language and Image Processing (ICALIP), 2010 International Conference on, 23-25 Nov. 2010 2010. 1056-1061.

UNDERWORLD, M. T. 2011. Mapping the Underworld [Online]. Available: http://www.mappingtheunderworld.ac.uk [Accessed 13/09 2015].

Ward, G., Hastie T Fau - Barry, S., Barry S Fau - Elith, J., Elith J Fau - Leathwick, J. R. & Leathwick, J. R. Presence-only data and the em algorithm.

Werman, M. & Keren, D. A novel Bayesian method for fitting parametric and non-parametric models to noisy data. Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., 1999 1999. 1-558 Vol. 2.

Yangming Li, S. L., Quanjun Song, Hai Liu, Meng, M.Q.-H. 2013. Fast and Robust Data Association Using Posterior Based Approximate Joint Compatibility Test. IEEE Transactions on Industrial Informatics, 331 - 339.

Dutta, R., Cohn, A.G. & Muggleton, J.M., 2013. 3D mapping of buried underworld infrastructure using dynamic Bayesian network based multi-sensory image data fusion. Journal of Applied Geophysics, 92, pp.8–19. Available at: http://www.sciencedirect.com/science/article/pii/S0926985113000311.

Muggleton, J.M., Brennan, M.J. & Gao, Y., 2011. Determining the location of buried plastic water pipes from measurements of ground surface vibration. Journal of Applied Geophysics, 75(1), pp.54–61. Available at: http://www.sciencedirect.com/science/article/pii/S0926985111001376.

Thomas, A.M. et al., 2009. Stakeholder needs for ground penetrating radar utility location. Journal of Applied Geophysics, 67(4), pp.345–351. Available at: http://www.sciencedirect.com/science/article/pii/S0926985108000840.