

# **Determining Urban Emotion using an Unsupervised Learning Approach: A Case Study around Majitar, East District, Sikkim**

Supriya Choudhury  
Sikkim Manipal University  
Dept. of CSE, SMIT  
Majitar, Sikkim

Mohan P. Pradhan  
Sikkim University  
Dept. of CA  
Gangtok, Sikkim

Pratikshya Sharma  
Sikkim Manipal University  
Dept. of CSE, SMIT  
Majitar, Sikkim

S. K. Kar  
Sikkim Manipal University  
Dept. of CSE, SMIT  
Majitar, Sikkim

## **ABSTRACT**

Perception and expectation of citizens is an important factor in urban settlement, planning and management. Hence, there is a need of a participatory citizen centric planning of urban settlement based on spatial data. These perception and expectation may be represented in terms of emotions. Determining Urban Emotions is an approach which can be used to map different types of emotions associated with urbanization. In the recent years, some new methods have been presented for the area of urban and spatial planning that resulted in a fundamental change of the issues and understanding of urban planning. Geographical information system acts as a key factor for analyzing urban emotions from various types of data. This paper presents the unsupervised learning approach for determining urban emotions using K-Means algorithm.

## **Keywords**

Urban Planning, Spatial planning, Smart city, Urban Emotions, K-Means algorithm

## **1. INTRODUCTION**

An Urban Emotion [2] is one of the emerging approaches that combine the concepts of spatial planning, geographic information systems, computer linguistics, sensor technology methods and real world data, where spatial planning considers all social and spatial structures within the city and helps in collecting various forms of data in context to the city [1]. It involves both the spatial and temporal patterns that help in research activities in identifying processes and to characterize special social-cultural movements and developments [1]. Geographic information system consists of two distinct disciplines geography and information system [1]. It is an information system designed to work with data that are referenced by spatial or geographical coordinates [1]. Computer linguistics is an interdisciplinary field which is concerned with the statistical and rule based modeling of natural language from a computational perspective. Real world data is an umbrella term used for different types of data that are collected in conventional randomized controlled trials [1]. It can be technical sensor data, crowd sourced data, human- sensor data or social data, etc [1] [3]. It can be used for decision making [1].

The main idea behind this approach is the involvement of people of a particular location into various planning processes [1]. Urban Emotion deals with different expectations of people regarding a particular location and what additional features can be added to the locality [1]. It explains the potentiality of integrating objectively quantifiable emotions in context of citizen participation [1]. Determination of Urban Emotion figures out the use of real world data [1].

Citizen's perception and urban space when linked together triggers an emotional reaction and creates its own atmosphere in the observer [1]. Urban emotion aims to understand how people's feelings get affected by features of the current environment, green spaces, air pollution, water pollution, noise pollution, affects of industrialization, land degradation, road condition, and other geographical factors [1].

Better urban planning approaches are needed to build a city into a smart city [1] [2] [5] [4]. Smart cities [6] are the cities that are able to operate in a sustainable, efficient and intelligent manner and require smart infrastructure with advanced sensing capabilities that extend beyond mere technical subtleties, thereby possibly benefitting architects and citizens of the cities [1]. It means smart citizens can make intelligent cities [1]. It relies on the idea that only citizens can make a city really intelligent [1]. It needs to be tackled both from technological view point and human centric view point that a city requires smart citizens to be intelligent themselves [1].

## **2. PROBLEM DEFINITION**

People's feelings and emotions generally changes with the geographical location. How people gain perception with context to the city is always been an issue in urban planning and management. The problems of considering subjective measurements and views provided by the citizens by involving them into planning processes represent great challenges for efficient urban planning. The wide range of the problems of extracting human emotion in context to the city may make good understanding of different expectation of the people. Assessing human emotions with relation to various geographical data is an important issue in urban planning. Urban Emotion adds a new information layer which will help in urban planning.

To solve this problem, we will be using different types of data of different facilities that are considered to be essential for urban planning and settlements. This case study is conducted for better understanding of the developments that took place so far in context to each facility, what are the different expectation of the people regarding each facility, how much people are emotionally attached to that location, how much knowledge do they have regarding Majitar location, what additional features and developments they want in that locality.

### 3. PROPOSED APPROACH

Urban emotions can be categorized by studying the demands of the individuals availing the following six facilities:

- (i) Educational Facility
- (ii) Entertainment Facility
- (iii) Health Facility
- (iv) Industrialization Facility
- (v) Shopping Mart Facility
- (vi) Transportation Facility

Urban emotion can be determined using the following approach:

Step1. Find and determine the types of emotions to be analyzed.

Step2. Building up questionnaires for emotions.

Step3. Determination and expression of emotions by the Sample Data Set.

Step4. Determining relationship between emotions by analysis of the expressions using a computational technique.

### 4. METHODOLOGY USED

The k-means algorithm is a type of unsupervised machine learning approach which is useful in performing unsupervised data classification. It is a non-hierarchical data clustering algorithm. The k-means algorithm belongs to a group of algorithms named as partitioning methods. The k-means algorithm initializes k clusters by performing arbitrary selection of one object for representing each of the clusters. The remaining objects are assigned to a cluster and we use a clustering criterion to calculate the cluster mean. These cluster means are further used as the new measured cluster point and each of the objects are reassigned to the cluster with which it has most similarity. The entire process continues until there is no longer possible of change when the clusters are recalculated.

K-means Algorithm:

1. First, select 'k' (number of) clusters arbitrarily.
2. Then, initialize cluster centers with those k clusters.
3. Next, start do loop

(a) Partition the data by assigning or by reassigning all the data objects to their closest and nearest cluster center using Euclidean distance.

(b) Compute the new cluster centers as a mean value of the objects in each cluster (that is, moves each cluster center to the mean of its assigned items) until no change in cluster center calculation or until convergence (change in cluster assignments is less than a threshold).

### 5. CASE STUDY AREA

The methodology is implemented in the Majitar. It is located at 27.1894°N 88.4978°E. It is a small village in the Indian state of Sikkim. The nearest towns to Majitar are Rangpo, which is about 4 km away and Singtam, which is 7 km away. Its location is about 200 metres (660 ft) above sea level which gives it a sub-tropical climate [11].

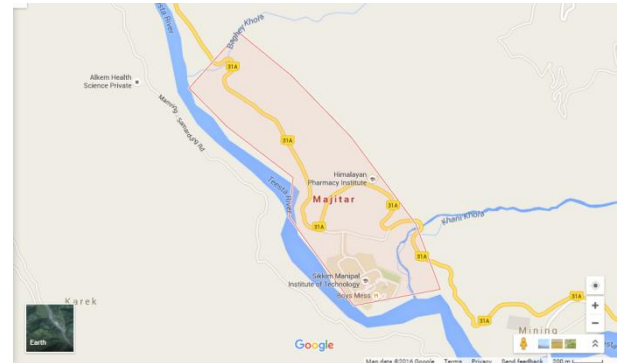


Fig 1: Map of Majitar (Source: Google Maps)



Fig 2: Map of Majitar (Source: Google Earth)

Majitar is largely populated by Nepalese people, Bhutia people, Marwari people, and Bengali people. Nepali language is the predominant language but some people speak Hindi too. Due to its geographic location and low population density, it has been considered to have less urban settlements and development [11].

### 6. RESULTS

For the implementation, a set of questionnaires for each facility was prepared. The number of questions varies from 10-12 numbers for each facility. The responses to each questions need to be provided in the range of 1-10 for specifying the quantities. Some questions require the responses in the form of nominal values such as worst, bad, average, good and better for better understanding of present condition and scenarios of each facility. And few questions required to be answered as yes or no. Then, the data was collected from 100 people of age group 15-60 residing in Majitar locality by conducting a survey [4]. A dataset of range 1 to 100 was prepared. Out of 100 people, 66 responses are provided by male participants and 34 are provided by female participants. For determination and expression of emotions by the sample set, we tried to analyze the responses provided by the people of that particular location. The implementation has been done with the help of IBM Statistics SPSS software. Initially, the number of clusters was arbitrarily considered and determined, that is, k=3 for each facilities respectively. In this

paper, the result obtained for first three facilities are shown as below:

### 6.1 Industrialization Facility Data

It consists of 10 numbers of questions. Q1 is about necessity of industrialization at Majitar. Q2 is for getting an opinion about whether the increase in industrialization is leading to environmental pollution at Majitar. Q3 and Q4 are for need of industrialization for economic growth and whether there is a need of more industries to be established at Majitar. Q5 is all about different types of industries that are supported by the people of Majitar. Q5 has 10 types of industries namely; Q51 refers to engineering and machinery industry, Q52 refer to tourism industry, Q53 refer to transportation industry, Q54 refer to chemical industry, Q55 refers to IT and ITES industry, Q56 refer to textile industry, Q57 refer to agro based industry, Q58 refers to food and beverage industry, Q59 refer to mineral based industry and Q510 refer forest based industry. Q6, Q7 and Q8 are for getting an opinion from the people of Majitar about whether development in industrialization will increase in employment in that location, whether it will help in increasing the real estate value of the land and should government provide more land for industrial development respectively. Q9 is about whether there is any development so far in this facility. Q10 is about present condition of this facility.

**Table 1. Initial Cluster Centers for Industrialization Facility Data**

	Initial Cluster Centers		
	Cluster		
	1	2	3
Q1	1	1	1
Q2	0	0	0
Q3	1	1	1
Q4	1	0	1
Q51	1	1	0
Q52	0	1	0
Q53	0	0	0
Q54	0	1	0
Q55	0	0	0
Q56	0	1	0
Q57	0	1	1
Q58	0	1	0
Q59	0	0	0
Q510	0	1	1
Q6	0	1	1
Q7	0	1	1
Q8	0	1	1
Q9	0	1	0
Q10	2	4	0
AGE	18	59	37

**Table 2. Iteration History for Industrialization Facility Data**

Iteration	Iteration History		
	Change in Cluster Centers		
	1	2	3
1	4.070	5.380	3.895
2	.099	.650	.245
3	.000	.000	.000

Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 19.261.

**Table 3. Final Cluster Centers for Industrialization Facility Data**

	Final Cluster Centers		
	Cluster		
	1	2	3
Q1	1	1	1
Q2	1	0	0
Q3	1	1	1
Q4	1	1	1
Q51	0	1	1
Q52	1	1	1
Q53	0	0	0
Q54	0	1	1
Q55	0	0	0
Q56	0	1	1
Q57	0	1	1
Q58	1	1	1
Q59	0	0	0
Q510	1	1	1
Q6	1	1	1
Q7	1	1	1
Q8	1	1	1
Q9	0	1	1
Q10	2	3	3
AGE	21	53	36

**Table 4. Distance between Final Cluster Centers for Industrialization Facility Data**

Cluster	Distances between Final Cluster Centers		
	1	2	3
1		31.663	14.079
2	31.663		17.598
3	14.079	17.598	

### 6.2 Shopping Mart Facility Data

It consists of 6 numbers of questions. Q1 is about availability of shopping mart at Majitar. Q2 gives the number of shopping mart available, and Q3 queries whether shopping mart are necessary at Majitar. Q4 is about availability of sufficient area for establishing more shopping marts at Majitar. Q5 queries about what more facilities should be there in shopping mart. Q5 has a list of ten facilities which are to be newly added to this facility, namely, Q51 is for having food court in that shopping mart; Q52 is for having a vegetable, fruit and fish market in that shopping mart; Q53 is for having a designer stores for clothing and shoes in it ; Q54 is for having stores for accessories and household necessities; Q55 is for having a playhouse for children in it; Q56 is for having a game parlor in the shopping mart; Q57 is for having a gym in it; Q58 is for having beauty salons for men and women; Q59 is for having some stores for stationary items and Q510 is for having multiplexes and PVRs in it. Q6 is about present condition of this facility.

**Table 5. Initial Cluster Centers for Shopping Mart Facility Data**

	Initial Cluster Centers		
	1	2	3
Q1	0	0	0
Q2	0	0	0
Q3	1	1	1
Q4	1	0	0
Q51	1	1	1
Q52	1	1	1
Q53	0	0	0
Q54	0	1	1
Q55	1	1	1
Q56	0	0	0
Q57	0	0	0
Q58	0	1	0
Q59	1	1	0
Q510	1	1	0
Q6	3	0	2
AGE	59	37	18

**Table 6. Iteration History for Shopping Mart Facility Data**

Iteration	Iteration History		
	Change in Cluster Centers		
	1	2	3
1	5.354	2.741	3.975
2	.649	.233	.098
3	.000	.000	.000

The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 19.183.

**Table 7. Final Cluster Centers for Shopping Mart Facility Data**

	Final Cluster Centers		
	1	2	3
Q1	0	0	0
Q2	0	0	0
Q3	1	1	1
Q4	1	1	1
Q51	1	1	1
Q52	1	1	1
Q53	0	0	1
Q54	0	1	1
Q55	1	1	1
Q56	0	0	1
Q57	0	0	1
Q58	0	0	1
Q59	1	1	1

Q510	1	1	1
Q6	2	2	1
AGE	53	36	21

**Table 8. Distance between Final Cluster Centers for Shopping Mart Facility Data**

Cluster	Distances between Final Cluster Centers		
	1	2	3
1		17.598	31.654
2	17.598		14.075
3	31.654	14.075	

### 6.3 Transportation Facility Data

It consists of 12 numbers of questions. Q1 and Q2 queries whether a participant owns a vehicle, if yes, then how many vehicles are owned. Q3 is about availability of public transport at Majitar. Q4 is for the need of three wheeler public transport at Majitar. Q5 is about availability of enough parking facility at Majitar. Q6 is about the present road condition. Q7 queries whether there is a frequent occurrence of road accidents. Q8 is for need of building up new speed breakers in order to avoid road accidents. Q9 is for the need of more attention to be drawn in road safety. Q10 is about availability of government funding regarding this facility. Q11 is about whether there is any development so far in this facility. Q12 is about present condition of this facility.

**Table 9. Initial Cluster Centers for Transportation Facility Data**

	Initial Cluster Centers		
	1	2	3
Q1	0	0	0
Q2	0	0	0
Q3	1	0	0
Q4	0	1	0
Q5	1	1	0
Q6	3	4	3
Q7	0	1	1
Q8	1	1	0
Q9	1	1	1
Q10	0	1	1
Q11	0	0	0
Q12	0	3	2
AGE	37	59	18

**Table 10. Iteration History for Transportation Facility Data**

Iteration	Iteration History		
	Change in Cluster Centers		
	1	2	3
1	3.218	5.320	3.739
2	.218	.651	.096
3	.000	.000	.000

The maximum absolute coordinate change for any center is .000. The current iteration is 3. The minimum distance between initial centers is 19.235.

**Table 11. Final Cluster Centers for Transportation Facility Data**

**Final Cluster Centers**

	Cluster		
	1	2	3
Q1	0	0	0
Q2	1	0	0
Q3	0	0	1
Q4	1	1	0
Q5	1	1	0
Q6	4	4	3
Q7	1	1	1
Q8	1	1	1
Q9	1	1	1
Q10	1	1	1
Q11	0	0	0
Q12	2	3	3
AGE	36	53	21

**Table 12. Distance between Final Cluster Centers for Transportation Facility Data**

**Distances between Final Cluster Centers**

Cluster	1	2	3
1		17.600	14.095
2	17.600		31.650
3	14.095	31.650	

During implementation, Euclidean Distance is used to assign or reassign all data objects to their closest and nearest cluster center. Euclidean Distance is the distance between point's p and q. It is actually the length of the line segment connecting them ( $\overline{pq}$ ). If  $C_1 = (p_1, p_2, p_3, \dots, p_n)$  and  $C_2 = (q_1, q_2, q_3, \dots, q_n)$  are two clusters. Then, the distance between the clusters can be given as:

$$d(p, q) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + \dots + (b_n - a_n)^2}$$

$$d(p, q) = \sqrt{\sum_{i=0}^n (q_i - p_i)^2}$$

In the tables given below are the results of Analysis of Variables (ANOVA) for each facility. The analysis for each variable is performed using the following parameters:

a) *df*: It stands for degrees of freedom and is equal to n-1.

b) *Mean Square*: It refers to an estimate of population variance based on the variability among a given set of measures. It is calculated by dividing the corresponding sum of squares by the degrees of freedom.

c) *Mean Square of Error*: It is obtained by dividing the sum of squares of the residual error by the degrees of freedom. It can be given as:

$$M.S.E = \frac{1}{n} \sum_{i=0}^n (\bar{Y}_i - Y_i)^2$$

Where,  $Mean = \frac{1}{n} \sum_{i=0}^n$  and  $Error = (\bar{Y}_i - Y_i)^2$

d) *F*: Dividing mean square by mean square error gives F, which follows the F-distribution.

e) *F-test*: It can be any of the statistical test in which the test statistic has an F-distribution under the null hypothesis. It is used when comparing statistical models that have been fitted to a dataset, in order to identify the model that best fits the population from which we sampled the data. It is sensitive to non-normality. The F-test in ANOVA is used to access whether the expected values of a quantitative variable within the several pre-defined groups differ from each other.

$$F = \frac{\text{Explained Variance (or between group variability)}}{\text{Unexplained Variance (or within group variability)}}$$

Where, Variance is the average squared distance between the mean and each data value.

$$\delta^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

The F-tests are used only for the descriptive purposes because the clusters have been chosen in order to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and so it cannot be interpreted as tests of the hypothesis that the cluster means are equal.

**Table 13. ANOVA table for Industrialization Facility Data**

**ANOVA**

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Q1	.304	2	.192	97	1.583	.211
Q2	2.243	2	.210	97	10.689	.000
Q3	.027	2	.124	97	.218	.804
Q4	.392	2	.237	97	1.651	.197
Q51	.883	2	.239	97	3.692	.028
Q52	.337	2	.214	97	1.580	.211
Q53	.674	2	.179	97	3.755	.027

Q54	2.027	2	.196	97	10.356	.000
Q55	.674	2	.179	97	3.755	.027
Q56	2.215	2	.209	97	10.575	.000
Q57	1.171	2	.230	97	5.093	.008
Q58	.418	2	.244	97	1.714	.186
Q59	.145	2	.121	97	1.197	.307
Q510	.284	2	.222	97	1.280	.283
Q6	.049	2	.048	97	1.013	.367
Q7	.042	2	.131	97	.319	.728
Q8	.224	2	.148	97	1.516	.225
Q9	1.528	2	.224	97	6.831	.002
Q10	9.287	2	1.571	97	5.910	.004
AGE	5747.064	2	9.473	97	606.692	.000

**Table 14. ANOVA table for Shopping Mart Facility Data**

**ANOVA**

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Q1	.056	2	.057	97	.985	.377
Q2	.353	2	2.006	97	.176	.839
Q3	.070	2	.057	97	1.235	.295
Q4	.340	2	.209	97	1.624	.202
Q51	.224	2	.104	97	2.153	.122
Q52	.381	2	.116	97	3.278	.042
Q53	2.803	2	.200	97	14.048	.000
Q54	1.347	2	.223	97	6.029	.003
Q55	.382	2	.209	97	1.831	.166
Q56	2.517	2	.206	97	12.230	.000
Q57	2.215	2	.209	97	10.575	.000
Q58	1.612	2	.221	97	7.300	.001
Q59	.747	2	.167	97	4.471	.014
Q510	.024	2	.176	97	.139	.871
Q6	6.908	2	1.600	97	4.318	.016
AGE	5747.064	2	9.473	97	606.692	.000

**Table 15. ANOVA table for Transportation Facility Data**

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Q1	.047	2	.108	97	.439	.646
Q2	.669	2	.618	97	1.083	.343
Q3	3.503	2	.156	97	22.496	.000
Q4	1.014	2	.236	97	4.299	.016
Q5	1.320	2	.230	97	5.738	.004
Q6	11.436	2	1.069	97	10.699	.000
Q7	.955	2	.215	97	4.444	.014
Q8	.917	2	.174	97	5.256	.007
Q9	.008	2	.138	97	.059	.943
Q10	.636	2	.190	97	3.347	.039
Q11	.760	2	.201	97	3.787	.026
Q12	.934	2	.816	97	1.145	.323
AGE	5747.064	2	9.473	97	606.692	.000

## 7. APPLICATIONS

There are various application areas of Urban Emotions. It helps in urban planning, settlements, development and safety. It is useful for traffic planning and people centric tourism. It may be used for assessing previous planning measures. It can be used to improve the quality of living of citizens.

## 8. CONCLUSION

Urban emotion is becoming one of the major areas of research that aims at quality planning and urban settlement prior to its implementation for the betterment of citizens and humanity at large. From the results of an unsupervised learning approach K-Means Clustering algorithm, it can be concluded that the facilities in which the featured attributes has less value of the exact significance level of analysis of variance (ANOVA) table will effect in urban development and those featured attributes are said to have the best response than the other attributes of each facilities. More development should be made to those featured attributes which has less exact significance level value. In future, better urban planning and involvement of smart citizens will help in creating and building a Smart City [4] which will lead to better sustainability.

## 9. REFERENCES

- [1] Supriya Choudhury, Mohan P. Pradhan, S. K. Kar, A Survey on Determining Urban Emotions using Geo-Data Classification: A Case Study around Majitar, East District, Sikkim, International Journal of Computer Applications, (0975 – 8887), Volume 135 – No.2, February, 2016, ISBN: 973-93-80891-05-5.
- [2] Peter Zeile, Bernd Resch, Linda Dorrzapf, Jan-Philipp Exner, Gunter Sagl, Anja Summa, Martin Sudmanns, Urban Emotions–Tools of Integrating People’s Perception into Urban Planning, Conference Proceedings REAL CORP 2015 Tagungsband, 5-7 May 2015, Ghent, Belgium. ISBN: 978-3-9503110-8-2 (CD-ROM); ISBN: 978-3-9503110-9-9 (Print).
- [3] Peter Zeile, Bernd Resch, Jan-Philipp Exner and Gunther Sagl, Urban Emotions Benefits and Risks in Using Human Sensory Assessment for the Extraction of Contextual Emotion Information in Urban Planning, Springer International Publishing, 2015.
- [4] Bernd Resch, Martin Sudmanns, Gunther Sagl, Anja Summa, Peter Zeile, and Jan-Philipp Exner, Crowd-sourcing Physiological Conditions and Subjective Emotions by Coupling Technical and Human Mobile Sensors, GI Forum – Journal for Geographic Information Science, 1-2015, Berlin, ISBN 978-3-87907-558-4, ISSN 2308-1708, doi:10.1553/giscience2015s514.
- [5] Gunther Sagl, Bernd Resch, and Thomas Blaschke, Contextual Sensing: Integrating Contextual Information with Human and Technical Geo-Sensor Information for Smart Cities, Open Access Sensors, 2015, 15, 17013-17035; doi: 10.3390/s150717013, ISSN 1424-8220.
- [6] Bernd Resch, Anja Summa, Gunther Sagl, Peter Zeile, Jan-Philipp Exner, Urban Emotions–Geo-Semantic Emotion Extraction from Technical Sensors, Human Sensors, Springer International Publishing, 2014.
- [7] Chrysaida-Aliki Papadopoulou and Maria Giaoutzi, Crowd-sourcing as a Tool for Knowledge Acquisition in Spatial Planning, Future Internet 2014, 6, 109-125; ISSN 1999-5903, doi:10.3390/fi6010109.
- [8] Benjamin S. Bergner, Jan-Philipp Exner, Martin Memmel, Rania Raslan, Dina Taha, Manar Talal, Peter Zeile, "Human Sensory Assessment Methods in Urban Planning – a Case Study in Alexandria", Conference Proceedings REAL CORP 2013, Tagungsband, 20-23 May 2013, Rome, Italy, ISBN: 978-3-9503110-4-4 (CD-ROM); ISBN: 978-3-9503110-5-1 (Print).

- [9] Bernd Resch, “People as Sensors and Collective Sensing-Contextual Observations Complementing Geo-Sensor Network Measurements”, Springer International Publishing, 2013.
- [10] Peter Zeile, Martin Memmel, Jan-Philipp Exner, “A New Urban Sensing and Monitoring Approach: Tagging the City with the RADAR SENSING App”, Reviewed Paper of Conference Proceedings REAL CORP 2012, Tagungsband, 14-16 May 2012, Schwechat, ISBN: 978-3-9503110-2-0 (CD-ROM); ISBN: 978-3-9503110-3-7 (Print).
- [11]About Majitar, <https://en.wikipedia.org/wiki/Majitar>