

ANALYSIS AND DATA MINING OF LEAD-ZINC ORE DATA

Vladimir Zanev, Stanislav Topalov, Veselin Christov

ABSTRACT. This paper presents the results of our data mining study of Pb-Zn (lead-zinc) ore assay records from a mine enterprise in Bulgaria. We examined the dataset, cleaned outliers, visualized the data, and created dataset statistics. A Pb-Zn cluster data mining model was created for segmentation and prediction of Pb-Zn ore assay data. The Pb-Zn cluster data model consists of five clusters and DMX queries. We analyzed the Pb-Zn cluster content, size, structure, and characteristics. The set of the DMX queries allows for browsing and managing the clusters, as well as predicting ore assay records. A testing and validation of the Pb-Zn cluster data mining model was developed in order to show its reasonable accuracy before being used in a production environment. The Pb-Zn cluster data mining model can be used for changes of the mine grinding and floatation processing parameters in almost real-time, which is important for the efficiency of the Pb-Zn ore beneficiation process.

1. Introduction. The types of minerals and their grade or concentration in ore have a major impact on the operation and control of the processing

ACM Computing Classification System (1998): H.2.8, H.3.3.

Key words: data analysis, data mining, clustering, prediction, Pb-Zn ore data.

technologies and production. The relatively low ore grade of refractory types increases the complexity of the extraction of metal and other products. Understanding the ore grades with its variability, classification, and prediction can be a very useful tool to improve and control the operation processes of metal production.

In our study we focus on statistical analysis and data mining of data from an underground Pb-Zn (lead-zinc) mine in Bulgaria. One of the main Pb-Zn deposits in Bulgaria is in the Madan-Rudozem region in the south-east part of the Rhodope Mountains. Mining and metallurgy in this area have a long history, since Roman times, and the ore deposit fields are well established and known. The Madan-Rudozem mine and enrichment facility exploits the north-west part of the Pb-Zn ore deposit field. For data analysis and visualization we have used Matlab [5] and for data mining, SQL Server 2012 [6, 7].

2. Data Analysis. The dataset of our study consists of 722 records of ore assays. They are organized in blocks by level elevation. A block has the relative shape of a spatial parallelepiped containing ore assays data with level specification. It is used in the data encoding to identify each record (see Figure 1). Each block contains different numbers of assay records. The levels and the blocks are as follows:

- Level 540 – block number 7, 9, and 11, encoded as 5407, 5409, and 5411
- Level 590 – block number 7, 9, and 11, encoded as 5907, 5909, and 5911
- Level 640 – block number 11, 13, 15, and 17, encoded as 6411, 6413, and 6417

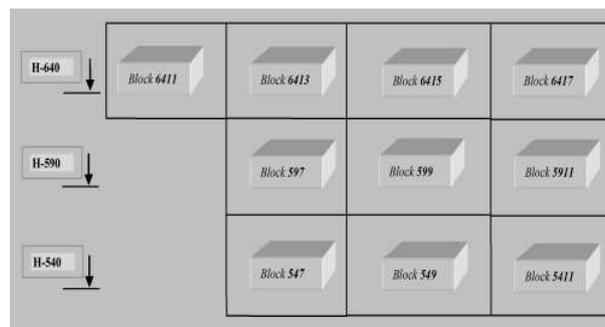


Fig. 1. Assay Blocks with Levels

The assay dataset records are made up of five attributes: X, Y, H (coordinates of the ore assay), Thickness (thickness of the ore vein), Pb (lead ore grade), and Zn (zinc ore grade). The X, Y, and H coordinates are the coordinates measured in meters with a local left-handed coordinates system, where the X axis indicates the assay distance north (positive), the Y axis indicates the assay distance east (positive) and H is the level of the ore assay. All attributes are real numbers and considered as continuous attributes. We have added two more calculated attributes: Pb/Zn and Pb/(Pb+Zn) ratios.

We examined the dataset for outliers. Outliers are abnormal data, real or erroneous, that can affect the quality of the data analysis results [4, 9]. We have identified three ore assay records with X or Y coordinates too far from the ore field coordinates and we removed these records from our data set. Below the ore assay locations are visualized in 3D and 2D space. The outliers are well visible.

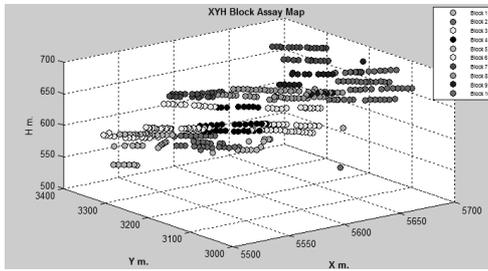


Fig. 2. 3D Block Assay Visualization

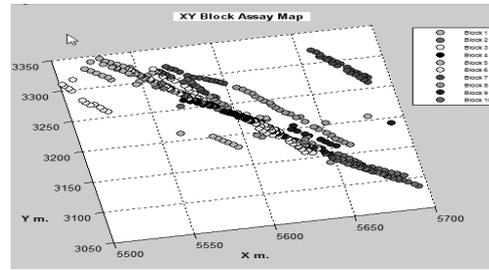


Fig. 3. 2D Block Assay Visualization

Between levels 640 and 690 a development of three small ore veins are established. They are not separate and independent ore veins but apophyses which are disconnected from the main vein and fade away by direction and face. The ore-bearing structure has a small but persistent thickness with a linear to slightly arcuate character.

The main basic statistics of the attributes of the Pb and Zn ore assays are given below.

Table 1. Assay Attributes Statistics

Attributes	Mean	StdDev	Min Value	Max Value
Ore Thickness	1.34	0.32	0.4	2.4
Pb	2.61	1.89	0.14	11.20
Zn	2.98	2.05	0.13	13.49
Pb/Zn	1.09	1.23	0.6	3.43
Pb/(Pb+Zn)	0.47	1.30	0.06	0.95

The ore vein thickness is relatively small but steady. Between the ore lead and zinc components, we established a positive moderate correlation $Pb = 0.5821 * Zn + 0.8976$, $r = 0.6258$. The lead and zinc mean grades are slightly below 3% but characterized with strong instability. The Pb/Zn ratio also shows instability and its estimation allows concluding a relative balance between the Pb and Zn grades. The Pb and Zn empirical distributions (see Figures 4 and 5 below) are with a significant left asymmetry and long tails from the right. The Weibull fading model seems to exhibit good fit to the Pb and Zn assay grades. We used Weibull probability distribution for interpolation with parameters $a = 2.9056$, $b = 1.4795$ for the Pb pdf (probability distribution function) and $a = 3.3246$, $b = 1.5325$ for the Zn pdf with over 90% interval of confidence.

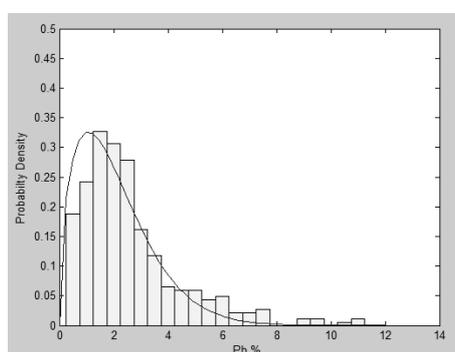


Fig. 4. Pb Probability Distribution

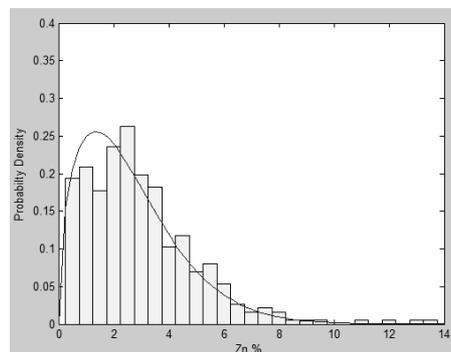


Fig. 5. Zn Probability Distribution

3. Cluster Data Mining of Pb-Zn Data. Data mining refers to the use of complex mathematical algorithms to perform tasks of classification, regression, segmentation, association, and sequence analysis, and to sift through detailed data to identify patterns and make predictions, correlations, and clusters within the data [4, 9]. A mine enterprise produces an enormous amount of data generated by different mine operations, systems, and components in a mine-wide information network. Data mining of mine enterprise data allows developing data mining models and use of these models to optimize mine operations and economies of scale. In the work of Coelho et al. [3], a neural network data mining model is used for assessment of petroleum wells operations; a data mining model is used to improve the shearer loader productivity in [1]; cluster data mining analysis of copper ore types with neural networks is reported in [2], and a design of production-oriented data mart for mine enterprise based on data mining is described in [8].

The difficulties to efficiently treat relatively low-grade, refractory-type ores require development of intelligent models to analyze and predict the ore content and grade. In our study we use clustering of the Pb-Zn assay data to create a data mining model for segmentation and prediction of Pb-Zn data.

3.1. Problem Definition, Data Preparation. The most common usage of data mining segmentation is to use clustering algorithms to detect the clusters in the data, label the clusters, use the clusters for analysis, report, and make predictions.

We have used the SQL Server 2012 Data Tools [7] to develop a cluster data mining model for segmentation and predictions of our Pb-Zn data. We used our cleaned from outliers Pb-Zn assay data set and created a SQL Server relational database with one table: PbZnData. The table's primary key is the composite key BlockNoPointNo (number of block and number of assay record in the block), X, Y, H assay coordinates, Distance (distance of the assay from the beginning of the coordinate system), Thickness (the value of the ore vein thickness), Pb, Zn grades, Pb/Zn, and Pb/(Pb+Zn) ratios.

3.2. Pb-Zn Cluster Data Mining Model. Using the PbZnData table as a source, we developed a data mining model and structure to apply the KMean non-scalable algorithm [7, 9]. The table primary key was selected as a key case of the mining structure; all other columns were set as Input type columns and the Pb and Zn columns as Predict type columns. We set the following cluster algorithm parameter: 5 clusters, drillthrough option, and 30% of training records. Five clusters were identified (see Figure 6), and we labeled them as High, Average-High, Average-Low, LowAvg-LowAvg, and Low, where the label implies the mean Pb-Zn grade for the cluster (see Table 2 below). For cluster analysis we have used the Cluster Viewer with its four tools – Cluster Diagram, Cluster Profile, Cluster Characteristics, and Cluster Discrimination [7].

Table 2. Pb-Zn Assay Cluster Mean and Size

Cluster Name	Pb Mean +/- StDev	Zn Mean +/- StDev	Size %
High	3.21 +/- 1.90	3.98 +/- 1.93	44
Average_High	1.89 +/- 1.28	3.89 +/- 2.51	11
Average_Low	2.40 +/- 1.66	1.74 +/- 0.96	33
LowAvg_LowAvg	1.98 +/- 0.88	1.56 +/- 0.81	7
Low	1.35 +/- 1.69	2.16 +/- 0.95	9

We use the Cluster Diagram to analyze the cluster content, size and the assay attribute values in the clusters. By default, the cluster shade represents the population of the cluster, and the shading of the line that connects one cluster

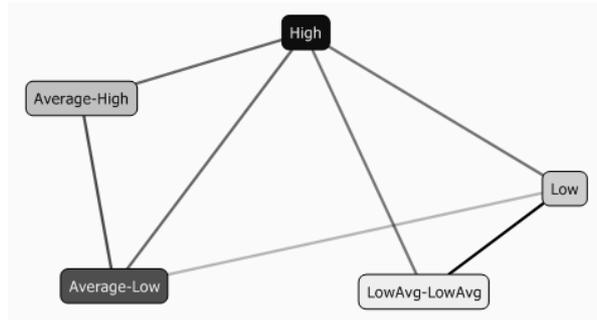


Fig. 6. Pb-Zn Assay Clusters

to another represents the strength of the similarity of the clusters. The darker shading represents a higher population and as the line becomes darker, the similarity of the links becomes stronger. We can select an attribute (shading variable) and interval of values (state) for the attribute to see where the assay records with these attribute values are clustered. For example, Pb grade with values > 5.15 are found mostly in the High cluster but some assay records with such values are found in the Average-Low cluster, or if we select the H attribute with values > 682 m., most of the records are grouped in the Low and LowAvg-LowAvg clusters.

The Cluster Profile provides an overall view of the clusters created. It calculates and displays for each cluster and each assay attribute the max, min, mean, standard deviation values and the size of the cluster. Table 2 shows the mean, standard deviation values, and the cluster size in percentage given by the Cluster Profile.

The Cluster Characteristics tool allows examining the characteristics that make up clusters in terms of attributes, their values, and order of importance, described by the probability that they appear in the cluster. In Table 3 below are summarized the Pb and Zn grade intervals ordered by probability of occurrence in the clusters.

We have used the Cluster Discrimination to determine the most important differences between clusters and the associated attributes with differences. For example, for the attribute H (the height of the assay) within interval (604 – 689) favors the High cluster and the interval (539-604) for H favors the Average-High cluster.

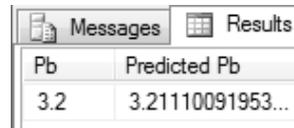
We are using the DMX (Data Mining Extension) language [7] to browse, manage, and analyze the cluster data mining model and make predictions against it. We created and executed DMX queries to extract cases and details about the cases in the clusters, filter the cluster content on different conditions, and

Table 3. Pb and Zn Cluster Grade Intervals

Cluster Name	Pb Grade Intervals (ordered by probability, low to high)	Zn Grade Intervals (ordered by probability, low to high)
High	(0.1-1.4]; (1.4-2.5]; (2.5-3.7]; (3.7-7.8]	(0.1-1.6]; (1.6-3.0]; (3.0-4.3]; (4.3-9.0]
Average_High	(2.5-3.7]; (3.7-7.8]; (0.1-1.4]; (1.4-2.5]	(0.1-1.6]; (1.6-3.0]; (3.0-4.3]; (4.3-9.0]
Average_Low	(0.1-1.4]; (3.7-7.8]; (2.5-3.7]; (1.4-2.5]	(3.0-4.3]; (0.1-1.6];(1.6-3.0]
LowAvg_LowAvg	(0.1-1.4]; (2.5-3.7]; (1.4-2.5]	(3.0-4.3]; (0.1-1.6];(1.6-3.0]
Low	(2.5-3.7]; (1.4-2.5]; (0.1-1.4]	(3.0-4.3]; (1.6-3.0]; (0.1-1.6]

get statistical summaries for the cluster data. An important part of our study is to develop and run DMX queries for prediction. We can apply the cluster data mining model to new data and make single or multiple predictions. Below you can see a singleton DMX query used to predict an existing case (an existing assay record) with good precision, which proves the validity of the mining model, and another DMX query for clustering of new assay data.

```
SELECT t.Pb, Predict([Pb]) AS [Predicted Pb] FROM [PbZnClusteringKmean]
NATURAL PREDICTION JOIN
(SELECT '6411 -18' AS [BlockNoPointNo], 5578.85 as X, 3302.7 AS Y, 644.31 as H,
1.2 as Thickness, 6515.1 as Distance, 3.20 as Pb, 5.5 as Zn,
0.366 as [Pb/(Pb+Zn)], 0.634 as [Zn/(Pb+Zn)], 0.578 as [Pb/Zn] ) AS t
```



Pb	Predicted Pb
3.2	3.21110091953...

Fig. 7. Pb and Predicted Pb grade for existing case

```
SELECT t.Pb, t.Zn, Cluster(), ClusterProbability() as [Cluster Probability]
FROM [PbZnClusteringKmean]
NATURAL PREDICTION JOIN
(SELECT 5078 as X, 3345 AS Y, 625 as H, 1.5 as Thickness, 2.25 as Pb,
3.15 as Zn) AS t
```

The variability of the Pb-Zn ore type requires different processing conditions in the grinding and flotation circuits. The Pb-Zn cluster data mining model allows early and fast ore type detection and appropriate changes to achieve stability of the ore processing.

Messages		Results	
Pb	Zn	\$CLUSTER	Cluster Probability
2.25	3.15	Average-Low	1

Fig. 8. Clustering of new assay data

3.3. Testing. Before using the Pb-Zn cluster data mining model in a production environment, we have to ensure the model is making predictions with a desired accuracy. The SQL Server 2012 Data Tools with its Mining Accuracy Chart tool allows testing and validation of the Pb-Zn cluster data mining model. The Lift Chart shows the score of the lift. The lift is calculated as the ratio of the actual prediction probability to the marginal probability in the test cases. The score of our model is 0.85 for Pb and 0.87 for Zn values, which is not very high.

We applied cross validation to evaluate how good our Pb-Zn cluster data mining model is in terms of three statistical measures: RSME (Root Mean Square Error), MAE (Mean Absolute Error) and LS (Log Score). For cross validation, we used our testing data separately for Pb and Zn attributes split equally in 10 folds. The RSME represents the average error of the predicted value when compared to the actual value. The MAE is the average error of the predicted value to the actual value. It is calculated by obtaining the absolute sum of errors, and finding the mean of those errors. The LS represents the ratio between two probabilities, converted to a logarithmic scale. A log score is similar to a percentage. The cross-validation table below shows the average and standard deviations of the three measures RSME, MAE, and LS. They show a reasonable accuracy of the Pb-Zn cluster data mining model.

Table 4. Cross Validation RSME, MAE, and LS measures

	RSME	MAE	Log Score
Pb	Average = 1.72 Std Deviation = 0.16	Average = 1.30 Std Deviation = 0.14	Average = 2.06 Std Deviation = 0.38
Zn	Average = 1.87 Std Deviation = 0.22	Average = 1.48 Std Deviation = 0.19	Average = -2.12 Std Deviation = 0.16

4. Conclusions. We developed a Pb-Zn cluster data mining model over the Pb-Zn assay raw dataset. As a first step for the data mining, we examined the dataset statistically, created data visualizations, and cleaned some outliers from the data set. We are using the Pb-Zn dataset to develop a Pb-Zn cluster data mining model. The Pb-Zn cluster data mining model creates segmentation of the

Pb-Zn dataset by splitting the assay records into five clusters labeled according to the mean Pb and Zn grade content. We studied the Pb-Zn cluster data mining model to analyze the cluster content, size, structure, and characteristics. We have used the DMX language to develop several data mining queries for browsing, managing and predicting assay records. The DMX queries enable prediction of Pb-Zn assay records and performing new Pb-Zn assay record segmentation. The Pb-Zn cluster data mining model allows to be used for changes of the mine grinding and floatation processing parameters in almost real-time which is very important for the efficiency of processing. The cross validation of the Pb-Zn data mining model shows a reasonable accuracy.

Our future plans include developing a new time series data mining model of Pb-Zn data. The time series data mining model will allow for making short-term predictions based on past Pb-Zn assay records. In order to achieve this, we are working with the mine enterprise representative to extend the data set with ore assay time stamps and increase the size of the dataset.

REFERENCES

- [1] BALABA B., I. YOUSEF, I. GUNAWAN. Utilisation of Data Mining in Mining Industry: Improvement of the Shearer Loader Productivity in Underground Mines. In: Proceedings of the IEEE 10th International Conference on Industrial Informatics, INDIN 2012, Beijing, July 2012, 1041–1046.
- [2] CHRISTOV V., ST. TOPALOV. Cluster Analysis of Copper Ore Types by Neural Networks in Aid of Mine Operations Planning. In: Proceedings of the 3-rd Balkan Mining Congress (BALKANMINE 2009), Izmir, Turkey, 1–3 October 2009, ISBN 978-9944-89-782-2, 1–7.
- [3] COELHO D., M. ROISENBERG, P. FILHO, C. JACINTO. Risk Assessment of Drilling and Completion Operations in Petroleum Wells Using Monte Carlo and Neural Network Approach. In : Proceedings of the 37th Conference on Winter Simulation WSC '05, December 2005, Orlando, FL, USA, doi: 10.1109/WSC.2005.1574466
- [4] HAN J., M. KAMBER, J. PEI. Data Mining: Concepts and Techniques. Third Edition, ISBN 978-0123814791, Morgan Kaufmann Publ., 2011.
- [5] HANSELMAN D., B. LITTLEFIELD. Mastering MATLAB. ISBN-10: 9780136013303, Prentice Hall Publ., 2011.

- [6] LARSON B. *Delivering Business Intelligence with Microsoft SQL Server 2012*. 3rd Edition, ISBN 978-0071759380, McGraw Hill, 2012.
- [7] MACLENNAN J., Z. TANG, B. CRIVAT. *Data Mining with MS SQL Server 2008*. ISBN 978-0470277744, Wiley Publ., 2009.
- [8] XINRUI L., M. HONGBIN, Z. HONGDI, R. FENGYU. The research of building production-oriented data mart for mine enterprises based on data mining. In: *Proceedings of the Sixth International Conference on Natural Computation (ICNC)*, Yantai, Shandong, China, 2010, 2186–2189.
- [9] WITTEN I., E. FRANK, M. HALL. *Data Mining. Practical Machine Learning Tools and Techniques*, 3rd Edition, ISBN 978-0123748560, Morgan Kaufman Publ., 2011.

Vladimir Zanev
Columbus State University
TSYS School of Computer Science
4225 University Ave
Columbus, GA 31907, USA
e-mail: zanev_vladimir@columbusstate.edu

Stanislav Topalov
University of Mining and Geology
Mine Surveying and Geodesy Department
Studentski Grad
1700 Sofia, Bulgaria
e-mail: stopalov@gmail.com

Veselin Christov
University of Mining and Geology
Computer Science Department
Studentski Grad
1700 Sofia, Bulgaria
e-mail: veso@mgu.bg

Received November 26, 2013
Final Accepted December 5, 2013