Automatic Trend Tracking Model for Coalbed Methane Production Forecast

Zhu Qingzhong¹, Du Haiwei¹, Hu Qiujia¹, Fan Bin¹, Wang Jigang¹, Yu Jiasheng¹, Wang Hongli²

¹PetroChina Huabei Oilfield Company, Hebei, China
²China University of Petroleum, Beijing Beijing, China
cyy_zqz@petrochina.com.cn, *15210870454@163.com

Abstract. Nowadays, huge quantities of coalbed methane (CMB) well production data have been saved in the fields’ database, which declares that the industry of CMB enters the age of Big Data. The traditional gas production data analysis methods, such as the decline curve and type curve, are not effective for complex production conditions. So, we proposed an automatic segmentation trend tracking Model for the CMB production forecast, which is based on the production data and is more robust for the complex production condition.

1. Introduction
Under the conditions of the original coalbed methane (CBM) reservoir, the methane is in the coal seam as an adsorbed state. Due to the unique characteristics of the adsorption state of CBM, the prerequisite for the production of CBM is the transformation from adsorption state to free state, which process is called desorption. Based on the theory of CBM adsorption, the production process of CBM is divided into “water drainage - pressure drop - methane desorption - methane diffusion - methane seepage” [1]. The essence of CBM well production is to reduce the bottom-hole flowing pressure through water drainage, reducing the pressure of the reservoir, so that the adsorbed coalbed methane can desorb, and provide the CBM production.

The production forecast for CBM well plays a critical role in the process of CBM development. At present, the methods for predicting the productivity of CBM wells mainly include numerical simulation method, Arps Decline curve method, and the typical curve method. Due to the complicated development process of coalbed methane and the low accuracy of coal seam property data, the prediction accuracy of reservoir numerical simulation technology is poor. In the production process of CBM wells, the production conditions are changeable, and the steady decline stage often does not appear in the later stage of production like conventional oil or gas reservoirs. The method of decline curve is only applicable to coalbed methane wells with a stable decline in the later stage. Geological factors and drainage control will affect the production of coalbed methane wells. Sometimes several wells in a single well group will also exhibit different production characteristics, which reduces the adaptability of the typical curve method.

Autoregressive integrated moving average model (ARIMA), a typical time-series analysis method, can extract both strong trend information and the randomness of the CBM wells production process. Given the influence of various factors on the productivity of CBM wells and the large difference between wells, the actual production data of coalbed methane wells in the Fanzhuang mining area of Fanzhuang, Shanxi, China, and establishes was used to build an ARIMA modeling method to predict coalbed
methane production. The time-series cross-validation technique was added to make the model achieve
the effect of automatic optimization and fitting to improve the accuracy of CBM prediction.

2. MATHeMATICAL MODEL
An ARIMA model is often noted ARIMA (p, d, q), which represents three parts: Autoregressive part,
Moving Average part, and Integrated part.[2]

2.1. Autoregressive Model
Time-series is said to follow an autoregressive model of order p if the current value of the series can be
expressed as a linear function of the previous values of the series plus a random shock term. The general
equation of an autoregressive model of order p, AR(p), can be written as:
\[ x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \epsilon_t \]  
(1)
Where:
\( \phi_1, \phi_2, \ldots, \phi_p \): the autoregressive model parameters.

2.2. Moving Average Model
The moving average (MA) model describes a time-series that is a linear function of the current and
previous random shocks (\( \epsilon \)). The random shocks are also called errors, residuals, or a white noise
process. A time-series \( x_t \) is said to be a moving average process of order q, MA(q), if
\[ x_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \]  
(2)
Where:
\( x_t \): the current value of time-series data;
\( \epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-q} \): the current and previous errors or random shocks;
\( \theta_1, \theta_2, \ldots, \theta_q \): the moving average model parameters.

2.3. Autoregressive Moving Average Model
While the AR and MA models can be used for many data sets, they are not adequate for some data, and
a more general set of models is needed. The autoregressive moving average (ARMA) models contain
both AR and MA processes. Alternatively, a time-series \( x_t \) is said to follow an autoregressive moving
average model of orders p and q, ARMA(p, q), if \( x_t \) satisfies the following difference equation:
\[ x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \cdots - \phi_p x_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \]  
(3)

2.4. Autoregressive Integrated Moving Average Model
To model a given time-series with the ARMA process, the series must be stationary. This means that
both the expected values of the series and its autocovariance function are independent of time. Also, the
series must have stabilized variance and constant mean.
Most time-series are nonstationary but some can be transformed into a stationary series by
differencing. This process is often used to remove the trend, seasonality, and periodic variations of the
series, thus rendering the nonstationary time-series stationary. The differenced time-series, \( W_t \), can then
be analyzed and modeled like any other stationary time-series. After modeling the differenced time-
series, the output series is transformed back to the original raw data, \( X_t \), by reversing the order of
differencing. Inclusion of differencing in the formulation of the ARMA model results in the ARIMA
model. An ARIMA model predicts a value in a response time-series as a linear combination of its past values, past errors, and current and past values of other time-series. The order of an ARIMA model is usually denoted by the notation ARIMA \((p, d, q)\), where \(p\) is the order of the autoregressive component, \(d\) the order of the differencing, and \(q\) the order of the moving-average process. Therefore, all models discussed previously are subsets of ARIMA models. Mathematically, the ARIMA model is written as:

\[
W_t = \mu + \frac{\theta(B)}{\phi(B)} \varepsilon_t
\]

Where:
\(W_t\): the response series or a difference of the response series;
\(\mu\): a constant or intercept;
\(B\): the backshift operator (i.e. \(B x_t = x_{t-1}\)),
\(\phi(B)\): the autoregressive operator, represented as a polynomial in the backshift operator, \(\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p\); 
\(\theta(B)\): the moving-average operator, represented as a polynomial in the backshift operator: \(\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q\);
\(\varepsilon_t\): the random error or shock.

### 3. DATA PREPROCESSING

#### 3.1. Data Source

The data used were obtained from PetroChina Huabei Oilfield Company. The typical CBM well 001 was in Fanzhuang District, Shanxi, China. Well 001’s production data curve was shown in Fig.1:

![Figure 1: The raw production data of well 001](image)

#### 3.2. Outliers Detection and Data Cleaning

Noise and erroneous data are generated due to people mistake and sensor fault in the production process, which will affect the performance of the ARIMA model mainly based on time-series trends. Therefore, before the production capacity prediction, the production data needs to be preprocessed, mainly including outlier identification and data cleaning.
the Holt-Winters method, which is a three-exponential smoothing algorithm, is often used as the outlier detection method for time-series data, which considers the trend, level, and seasonal of the data. The common CBM production data is not seasonal, so double exponential smoothing can be used. The formula for double exponential smoothing is as follows:

\[
\ell_t = \alpha x_t + (1 - \alpha) (\ell_{t-1} + b_{t-1})
\]
\[
b_t = \beta (\ell_t - \ell_{t-1}) + (1 - \beta) b_{t-1}
\]
\[
x_{t+1} = \ell_t + b_t
\]

Where:
- \(\ell_t\): The level component of the data;
- \(b_t\): The trend component of the data;
- \(\alpha\): A weight for smoothing moving average value;
- \(\beta\): A weight for exponential smoothing.

The red points in Fig.2 are the outliers detected by the double exponential smoothing method. The red dotted lines are the confidence intervals of 95%.

![Figure 2: Outliers detected by double exponential smoothing method](image)

The ARIMA model mainly uses gas production data, so we can use the water production, flow pressure, and casing pressure data for the outliers of the gas production data. If outliers are generated at the same time point, we use the moving average and the maximum likelihood estimation method for interpolation cleaning. The cleaned data is shown in Fig.3.
4. Modeling Methodology

4.1. Model Identification
The manual process of using ARIMA to forecast involves five stages: visualizing the time-series; stationarizing the series, plotting autocorrelation and partial autocorrelation (ACF/PACF) charts and find optimal parameters; building the ARIMA model and making predictions [3].

4.1.1. Visualizing the time-series
Firstly, by visualizing the time-series data, it can be determined whether the data used has an overall trend, whether there is seasonality and which model to use. By observing the daily gas production data of the coalbed methane well in Fig.4 (a), it is found that some random factors such as human factors and geological factors make the trend of daily gas production not obvious, such as the impact of secondary fracturing and well shut. And the cumulative gas production in Fig.4 (b) shows a trend. At the same time, we can observe that there is no seasonality. The daily gas production is the first-order difference of the cumulative gas production. For the ARIMA model, the data stability needs to be increased by difference, so the use of the two has the same meaning.
4.1.2. Stationarizing the data

It is important for building the ARIMA model that the time-series data is stationary, which means the mean, variance, and the covariance of the $t^{th}$ term and $(t+i)^{th}$ term should not be a function of time. Dickey-Fuller test was used to check the stationarity of the cumulative production data, and Fig.5 shows that the cumulative production data is not stationary.

![Figure 5](image)

**Figure 5** The rolling mean and standard deviation of the cumulative production data

To make the data stationary is through taking the difference of the data. The first order of cumulative rate is the daily rate, which is not stationary either, so more order difference was taken. Fig.6 (c) shows that the second order of the data (the first order of the daily rate data) is stationary enough.

![Figure 6](image)

**Figure 6** The differences of production data

4.1.3. Estimating optimal parameters

First-order difference and second-order difference of the daily gas data were used to plot the ACF and PACF charts. Fig.7 shows the ACF plot of the first order difference of daily production data. A rapid decrease in the ACF plot indicates that the series is stationary. The partial autocorrelation function measures the degree of association between time-series observations when the effect of other time lags on the response time-series is held constant. The PACF is an extension of ACF and is used to examine serial dependencies for individual lags. Fig.8 shows the ACF plot of the first order difference of daily production data. However, the partial autocorrelation function has more significant lags.

Identifying the ARIMA ($p$, $d$, $q$) model:

- $d$ the order of integration, which means the number of differences needed to make the series stationary, is 1;
\( p \) is that the current series values depend on its previous values with some lag. From the PACF chart in Fig.7, it is 0, which is the biggest significant lag:

\( q \) refers to the current error depends on the previous with some lag. From the ACF chart in Fig.7, it is also 0 with the same logic.

![Figure 7](image)

**Figure 7** Autocorrelation plots of the first difference of daily production data

![Figure 8](image)

**Figure 8** Autocorrelation plots of the second difference of daily production data

### 4.1.4 Building the ARIMA Model

The estimation of optimal parameters of ARIMA is very subjective, so we used different combinations of parameters to estimate and choose the optimal. The parameters were estimated using function minimization procedures so that the sum of squared residuals is minimized.

The result and accuracy of the model will be discussed in the next sector.

### 4.2 Automatic Model Identification

To reduce the subjectivity of manual judgment and improve the accuracy of the model, it can estimate model parameters automatically. Firstly, the squared residuals were chosen as the loss function for the task, then a cross-validation method was used to evaluate the given model parameters, calculate the gradient. The traditional cross-validation method will lose the dependency of time-series data, so we use a time-series cross-validation method. Training model on a small segment of production data from the beginning to some time, then making predictions for the next \( t \) steps, and calculate an error. Then, we expand the training sample to \( t + n \) value, make predictions from \( t + n \) until \( t + 2n \), and move the test segment of production data until the last available date. As shown in Fig.9, we have as many folds as \( n \) will fit between the initial training sample and the last observation.
5. Results and discussion

Fig. 10 presents the raw production data and the linear interpolation by the ARIMA model. A very good match can be observed, making the time-series very representative of the original production data.

The curve shown in Fig. 11 was used to compare the ARPS decline curve method with the ARIMA model. After comparing the two models, predictions were made for 160 days. The accuracy of ARIMA was 98.9%, and of Decline Curve was just 68.03% respectively. ARIMA (0,1,0) performs better than Decline Curve Analysis. More importantly, the ARIMA predicted the right trend.

6. Conclusion

We developed ARIMA models for CBM wells in Fanzhuang district, Shanxi, China, to predict CBM production. For its high accuracy in a short time forecast, it can be used to guide the operation.

The ARIMA model not only can extract the overall trend information of the whole production but also obtain some non-pure random information with relevant properties.

For the production data with two peaks like the Well 001 well in the Fanzhuang district, it is more suitable to use the ARIMA model for prediction by increasing the periodic method.
Acknowledgment
This research was conducted under the financial support of the National Science and Technology Major Project of China (Grant No. 2017ZX05064004).

References
[1] Zhang Sui'an, Qin Peng. 2015. Study on interference well test technology between coalbed methane wells. J. Coal Science and Technology. 43(12):75-79. DOI=10.13199/j.cnki.cst.2015.12.016.
[2] Olorinu, O., & Sulaimon, A. A. 2014,. Application of Time-series Analysis to Predict Reservoir Production Performance. J. Society of Petroleum Engineers. DOI= 10.2118/172395-MS
[3] Al-Fattah, S. M. (2005, January 1). Time-series Modeling for U.S. Natural Gas Forecasting. International Petroleum Technology Conference. DOI=10.2523/IPTC-10592-MS
[4] Mohammad Fuad, I. I., Rosli, L. H., & Husni, H. 2019. Application of Hybrid ARIMA – Decision Tree Algorithm in Reservoir Pressure Surveillance. Society of Petroleum Engineers. DOI=10.2118/197762-MS
[5] I. Akkurt, A. Elkhayat, The effect of barite proportion on neutron and gamma-ray shielding, Annals of Nuclear Energy 51 (2013) 5-9. DOI=10.1016/j.anucene.2012.08.026
[6] Laboratory Signals, Systems. Signal Processing Deployment in Power Quality Disturbance Detection and Classification. International Conference on Computational and Experimental Science and Engineering (2016). DOI= 10.12693/APhysPolA.132.415
[7] Z. Cömerta, A.F. Kocamaz. Comparison of Machine Learning Techniques for Fetal Heart Rate Classification. International Conference on Computational and Experimental Science and Engineering (2016). DOI= 10.12693/APhysPolA.132.451
[8] A.S. Yükselä, Ş. F. Çankaya. Design of a Machine Learning Based Predictive Analytics System for Spam Problem. International Conference on Computational and Experimental Science and Engineering (2016). DOI= 10.12693/APhysPolA.132.500
[9] M. Djendel, O. Allaoui. Characterization of Alumina–Titania Coatings Produced by Atmospheric Plasma Spraying on 304 SS Steel. International Conference on Computational and Experimental Science and Engineering (2016). DOI= 10.12693/APhysPolA.132.538
[10] O. Ipek, B. Gürel. Numerical Investigation on Hydrodynamic Combustion and NOx Emission Behavior in 8 MW Circulating Fluidized Bed. International Conference on Computational and Experimental Science and Engineering (2016). DOI= 10.12693/APhysPolA.132.553