Energy Policy 82 (2015) 48–61

Contents lists available at ScienceDirect

Energy Policy

journal homepage: www.elsevier.com/locate/enpol

Modeling oil production based on symbolic regression

Guangfei Yang [a], Xianneng Li [b], Jianliang Wang [c], Lian Lian [d,*], Tieju Ma [e]

[a] School of Management Science and Engineering, Dalian University of Technology, Dalian, China
[b] Graduate School of Information, Production, and Systems, Waseda University, Kitakyushu, Japan
[c] School of Business Administration, China University of Petroleum, Beijing, China
[d] School of Transportation and Logistics, Dalian University of Technology, Dalian, China
[e] School of Business, East China University of Science and Technology, Shanghai, China

HIGHLIGHTS

- A data-driven approach has been shown to be effective at modeling the oil production.
- The Hubbert model could be discovered automatically from data.
- The peak of world oil production is predicted to appear in 2021.
- The decline rate after peak is half of the increase rate before peak.
- Oil production projected to decline 4% post-peak.

ARTICLE INFO

Article history:
Received 1 September 2014
Received in revised form
15 January 2015
Accepted 18 February 2015

Keywords:
Oil production
Hubbert theory
Symbolic regression

ABSTRACT

Numerous models have been proposed to forecast the future trends of oil production and almost all of them are based on some predefined assumptions with various uncertainties. In this study, we propose a novel data-driven approach that uses symbolic regression to model oil production. We validate our approach on both synthetic and real data, and the results prove that symbolic regression could effectively identify the true models beneath the oil production data and also make reliable predictions. Symbolic regression indicates that world oil production will peak in 2021, which broadly agrees with other techniques used by researchers. Our results also show that the rate of decline after the peak is almost half the rate of increase before the peak, and it takes nearly 12 years to drop 4% from the peak. These predictions are more optimistic than those in several other reports, and the smoother decline will provide the world, especially the developing countries, with more time to orchestrate mitigation plans.

1. Introduction

Oil is crucial for widespread transporting goods and people which is a hallmark of our modern civilization; accounting for 31.4% of total primary energy consumption and more than 90% of road-transport energy demand (IEA, 2013), as such it is important to model oil production into the future (IEA, 2013; EIA, 2013).

The methodology behind the modeling of oil production plays the core role in attempts to solve this problem, and among the many areas of discussion on this topic, quantitative models, which are usually built on mathematical formulae, are an important branch to study. Quantitative models, which date back to the early 19th century, have been built for measuring oil exhaustion (Day, 1909). Since then, more variations and improvements have emerged, which are usually classified into three approaches: (1) curve-fitting models; (2) system simulation; and (3) economic models (UKERC, 2009). The latter two approaches usually consider many related factors that affect the oil production to improve the models, such as oil price, demand, investment, etc., however, they are much more complicated to realize than the first one due to their reliance on a detailed description of the target area as well as the need for collection of a large amount of data. This restricts the applicability and effectiveness of the second and third approaches. The curve-fitting approach is relatively easier to implement, which makes it the widely accepted model to predict oil production (UKERC, 2009).

Among the various curve-fitting approaches, the Hubbert model is recognized as the most representative one. It was developed byHubbert (1956, 1982) and was successful in predicting oil production in lower 48 states of the USA. After Campbell and Laherrere's (1998) paper, Hubbert curves and the peak oil theory

* Corresponding author.
E-mail addresses: gfyang@dlut.edu.cn (G. Yang), xianneng.li@gmail.com (X. Li), wangjianliang305@163.com (J. Wang), lian.lian@dlut.edu.cn (L. Lian), tjma@ecust.edu.cn (T. Ma).

http://dx.doi.org/10.1016/j.enpol.2015.02.016
0301-4215/

took off with many extensions being based on it or a similar bell-shaped curve to predict future oil production (Szklo et al., 2007; Wang et al., 2011). This branch could be further classified into two classes: (1) symmetric models, like the Hubbert model and the Gaussian model; and (2) asymmetric models, like the Gompertz model. Each of these has a diverse impact on prediction performance (Sorrell et al., 2010). There was a comprehensive study to determine which model fitted best (Brandt, 2007) for 139 oil producing regions that are sub-national, national, and multi-national in scale. This empirical study proved that there is no model that is permanently superior to the others and it depends on a human expert's experience or judgment to choose the best model for a specific target area under study. Even if an appropriate model is selected, there will still be some parameters to be estimated, such as ultimately recoverable resources (URR) (Owen et al., 2010; Hirsch, 2005; Chapman, 2014).

The above discussions call for a more reliable and flexible method to model oil production to assist energy policy-making. In this study, we propose a novel method that is based on an evolutionary approach to symbolic regression. Symbolic regression was developed based on a popular evolutionary algorithm, genetic programming (Koza, 1992). There are various successful applications of symbolic regression and one of the representative works is the discovery of natural physical laws from data, such as nonlinear energy conservation laws and Newtonian force laws (Schmidt and Lipson, 2009). Many other interesting applications have emerged recently in different fields, including astronomy (Graham et al., 2013), biology (Sahakyan and Vendruscolo, 2013), and medicine (Yoshihara et al., 2013). We are inspired by the discovery of physical laws from data; accordingly, we attempt to find the underlying rules from oil production data. That is, we let the computer search the data and automatically discover the law-like models. The advantage is that it is not necessary to give assumptions or predefine the possible structures of models, which reduces the controversy and relieves the burden to build reliable models.

The purpose of our approach is to improve the curve-fitting research. The previous curve-fitting approaches usually assume a model and some parameters such as URR, which are not necessary in our approach. There are many other related factors that affect the oil production, and they could increase the precision of forecasting in some areas. However, they usually require a lot of data to collect and may not be suitable for some other areas, especially the larger regions, due to the lack of enough data. The curve-fitting approach usually does not consider so many factors, and the discussion of integration of other related factors is beyond the scope of this paper.

The rest of this paper is organized as follows. In Section 2, we briefly introduce the technical background of symbolic regression. In Section 3, we examine our approach on synthetic and real data to validate its modeling ability and predictability, and then apply our approach to predict the world oil production and analyze the features of the coming peak. In Section 4, we give some more discussions of our approach. In Section 5 we conclude the paper and discuss some policy implications.

2. Method

Genetic programming (GP) (Koza, 1992) is one of the classic evolutionary algorithms inspired by the Darwinian theory of evolution and is a successful variant of genetic algorithms (GA) (Holland, 1975). The evolutionary algorithms usually consists of three fundamental elements:

1. A randomly generated population of individuals that represent the candidate solutions.

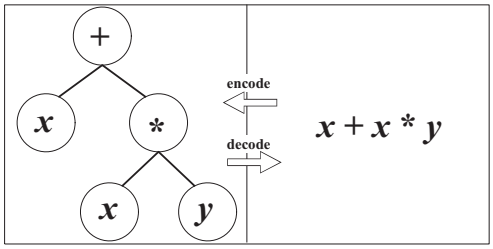

Fig. 1. A GP individual in tree structure.

2. Some genetic operators to evolve the individuals, of which the crossover (exchanging part of two individuals) and mutation (changing a small portion of one individual) are the common ones.
3. A fitness evaluation method to guide the direction of evolution, where the individual with higher fitness values will have a higher probability of surviving in the following evolution generations.

GP was designed originally to evolve the computer programs. The typical representation of GP is a tree structure, which could be evaluated in a recursive way, as shown in Fig. 1. There are generally two types of nodes: leaf nodes (or terminal nodes) and internal nodes (or functional nodes), where leaf nodes represent terminal symbols, such as variables and constants and internal nodes represent nonterminals, such as functions and operators. The overall flowchart of GP is illustrated in Fig. 2, and the crossover and mutation operators are explained in Figs. 3 and 4.

Symbolic regression is to automatically search for mathematical equations via GP from the high-dimensional space of a finite set of input data samples without any a priori expert's domain knowledge or pre-specified underlying regression structures like linear regression or nonlinear regression. Because of this unique advantage, symbolic regression has received more and more attention in various fields and many successful applications have emerged recently. Based on symbolic regression, Schmidt and Lipson (2009) developed an algorithm automatically searching motion-tracking data captured from various physical systems, and without any prior knowledge about physics, the algorithm discovered Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation. Vladislavleva et al. (2013) proposed an approach for energy prediction based on weather data and analyzed the important parameters as well as their correlation on the energy output. Can and Heavey (2011) developed metamodels to predict throughput rates in a common industrial system, without any prior assumptions on the structure of the metamodels. Kotanchek et al. (2010) detected outliers and extracted significant features from the country data to identify records that are systematically under- or over-predicted. Manson (2005) modeled decision making in the context of human–environment relationships, contributing to methodological innovations in multi-criteria evaluation and modeling of coupled human–environment systems. Khu et al. (2001) applied symbolic regression to real-time runoff forecasting for the Orgeval catchment in France. Yang et al. (2009) searched for optimized ranking equations to rank association rules by considering both objective and subjective information between the rules and keywords.

The procedures for symbolic regression via genetic programming are described in Algorithm 1, with a set of data samples $\mathcal{D} = \{(d_1, t_1), (d_2, t_2), \ldots, (d_n, t_n)\}$, where one data sample d_i ($i \in [1, n]$) consists of m dimensional variables; i.e., $d_i = \{d_{i1}, d_{i1}, \ldots, d_{im}\}$, and t_i ($i \in [1, n]$) is the target variable of d_i. The maximum number of generations of evolution is $\mathcal{L}$ and the output formula set is $\mathcal{F} = \{f_1, f_2, \ldots, f_p\}$, where f_j ($j \in [1, p]$) is one equation found by symbolic regression. The set of individuals in