CIT. Journal of Computing and Information Technology, Vol. 29, No. 3, September 2021, 151–165
doi: 10.20532/cit.2021.1005404

Application of Big Data Analysis to Agricultural Production, Agricultural Product Marketing and Influencing Factors in Intelligent Agriculture

Jianfeng Cheng

Economics and Management School, Wuhan University, Wuhan, China

Agricultural Internet of things (AIoT) promotes the modernization of traditional agricultural production and marketing model. However, the existing time series prediction methods for agricultural production and agricultural product (AP) marketing cannot adapt well to most real-world scenarios, failing to realize multistep forecast of production and AP marketing data. To solve the problem, this paper explores the big data analysis of agricultural production, AP marketing, and influencing factors in intelligent agriculture. To realize long-, and short-term predictions, a small-sample time series model was set up for AIoT production, and a big-sample time series model was constructed for AP marketing. The data fusion algorithm based on Kalman filter (KF) was adopted to fuse the massive multi-source AP marketing data. The proposed strategy was proved valid through experiments.

ACM CCS (2012) Classification: Applied computing → Computers in other domains → Agriculture

Keywords: agricultural production, agricultural product (AP) marketing, intelligent agriculture, big data analysis

1. Introduction

As a pillar of agricultural informatization, agricultural Internet of things (AIoT) has been extensively applied in planting, gardening, farming, logistics, and sales, and promotes the modernization of traditional agricultural production and marketing model [1–5]. The application of the Internet of things (IoT) to agricultural production and agricultural product (AP) sales leads to intelligent agriculture, which is smarter than traditional agriculture, bringing lots of convenience to our work and life. In intelligent agriculture, the agricultural production and AP sales are controlled by sensors and data crawlers, via mobile or computer platforms [6–8]. With the elapse of time, both the production data collected by sensors and online sales data gathered by crawlers will grow exponentially. The drastic increase of these time series data raises stricter requirements on the performance of intelligent agriculture management systems.

Researchers, engineers, and scholars around the world have been actively exploring the data collection, storage, and application of AIoT, and achieved quite a remarkable progress [9–11]. Ananthi *et al.* [12] provided an embedded system for soil monitoring and irrigation to reduce manual monitoring of farmlands, and acquired information via mobile apps to help farmers effectively boost agricultural yield. Matsumoto *et al.* [13] carried out numerical simulation based on the farmland information collected from the IoT, and realized advanced commercial evaluation of the inventory shortage and crop loss induced by the uncertainty of harvested crops. Whereas most intelligent agriculture systems focus on monitoring, Wongpatikaseree *et al.* [14] introduced the IoT to detect environmental data in intelligent farms, using multiple sensors, and proposed a traceable system that summarizes and displays the observations from intelligent farms. Before purchasing APs, a client can scan

the quick response (QR) code with a mobile app, and access various information on planting process and quality. Based on IoT sensors, Lee *et al.* [15] developed an agricultural production prediction and decision support system, and supported the sowing process by selling APs to consumers. Wang and Yang [16] integrated AIoT with sensor technology and agricultural big data, and achieved both growth environment and full lifecycle management (including growth, production, processing, distribution, and sales). Considering the features of modern agriculture and AP logistics, Mo [17] improved the traditional AP logistics model, presented a reference model for IoT-based AP supply chain, and analyzed the merits of the model. Harshani *et al.* [18] introduced IoT to monitor soil parameters like pH, soil temperature, and humidity, thereby improving crop productivity.

Owing to nonlinear correlations between samples, the existing time series prediction methods for agricultural production and AP marketing cannot adapt well to most real-world scenarios, failing to realize multistep forecast of production and AP marketing data. By virtue of the functional approximation ability of neural networks under big sample condition, this paper integrates deep neural network (DNN), which is known for its high prediction accuracy, to big data analysis on agricultural production and AP marketing. The main contents of this work is thus as follows: (1) the overall architecture of small-sample time series model for AIoT production was established to realize short-term prediction; (2) the overall architecture of time series data processing model for AIoT sales was created, and the massive multi-source AP marketing data were merged by a data fusion algorithm based on Kalman filter (KF); (3) a big-sample time series model was constructed for AP marketing to realize long-term prediction. The proposed strategy was proved valid through experiments.

2. Small-Sample Time Series Model for IoT Production and Short-Term Prediction

Thanks to the burgeoning of IoT and information technology, farmers can query the monitoring data of agricultural sensors on software platforms at anytime, and anywhere. This greatly facilitates agricultural production activities, and helps to reasonably allocate agricultural production resources, reduce agricultural production cost, and improve AP quality.

Focusing on agricultural production, this paper models the historical time series data, influencing factors, and their relationships with the time series data to be predicted. The influencing factors include soil, climate, and water source, to name but a few. Figure 1 shows the workflow of agricultural sensors during the monitoring of agricultural production.

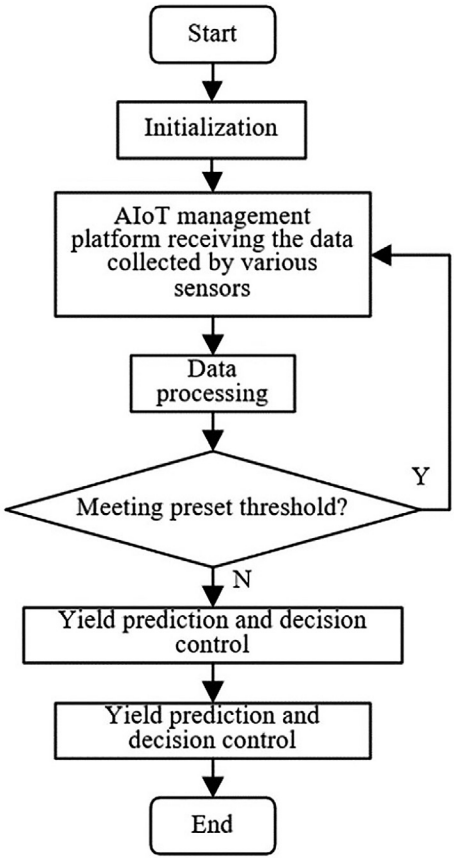


Figure 1. Workflow of agricultural sensors.

The signals collected by the sensors are amplified, and subjected to analog/digital (AD) conversion, before being transmitted to the controller of the AIoT management platform. The controller will process the received data, and judge if they meet the preset threshold. If the data are greater than the preset threshold, the data will be discarded, and new data will be received. If the data are smaller than the preset threshold, the platform will predict the yield,

and make control decisions, while adjusting the parameters of agricultural production activities.

This paper only predicts short-term time series, aiming to reduce the number of parameters in the prediction model, and improve the real-timeliness of prediction data. For a given time series $\{a(h)\}^{\varphi}_{h=1}$ of agricultural sensor monitored data at the current period h, the prediction model $G(*)$ can be established as:

$$a^*(h+1) = G\left(a\left(h - M^B : M^B\right),\ d(h+1)\right) \quad (1)$$

Let $\beta(h+1) = [a(h - M^B : M^B)\ d(h+1)]$ be the independent variable of the prediction model, which covers both the intrinsic attributes of agricultural production activities and the external attributes; $a^*(h+1)$ be the dependent variable characterizing the estimate of the time series a(h+1) in the subsequent period; $a(h - M^B : M^B) = ([a(h - M^B + 1), a(h - M^B), ..., a(h)] \in \mathbb{R}^{MB}$ be the intrinsic attributes of agricultural production activities, which characterize the M^B historical measurements of different agricultural sensors before period h; $d(h+1) = [d(h+1)1, d(h+1)2, ..., d(h+1)_{MI}\} \in \mathbb{R}^{MI}$ be the attribute vector of the attributes of agricultural production activities, which are composed of influencing factors at period $h+1$. Through functional fitting of the known data monitored by agricultural sensors, $G(*)$ can be obtained by solving a least squares (LS) problem. Based on a set $P = (\beta(h_m), a(h_m))\ M_{m=1}$ of M training samples, the solving process can be described as:

$$\begin{aligned} G &= \arg\min_{g \in G} Loss(C) \\ &= \arg\min_{g \in G} \sum_{C} loss\left[a(h_m), g\left(\beta(h_m)\right)\right] \end{aligned} \quad (2)$$

where $Loss(*)$ and $loss(*)$ are loss functions. The loss functions based on mean squared error (MSE) and absolute value can be respectively constructed as:

$$\begin{aligned} &loss\left[a(h_m), g\left(\beta(h_m)\right)\right] \\ &= \\ &\left[a(h_m), g\left(\beta(h_m)\right)\right]^2 \end{aligned} \quad (3)$$

$$\begin{aligned} &loss\left[a(h_m), g\left(\beta(h_m)\right)\right] \\ &= \\ &\left|a(h_m), g\left(\beta(h_m)\right)\right| \end{aligned} \quad (4)$$

To improve the prediction accuracy of the time series on a small dataset of the short-term monitoring on agricultural production activities, this paper proposes an additive expression to mine the nonlinear relationship between influencing factors and time series data. To reflect the degree of changes to the time series induced by external factors, the variable component of the time series in period h is denoted as $WS(h)$. To characterize the intrinsic structure of the time series, the intrinsic stationary component of the time series in period h is denoted as $TS(h)$. The proposed additive expression can decompose the time series $\{a(h)\}$ into two parts, which respectively consist of stationary features, and non-stationary features:

$$a(h) = WS(h) + TS(h) \quad (5)$$

Different from the intrinsic stationary component, the variable component has a complex, nonlinear relationship with influencing factors, and cannot be modeled through classic time series analysis. To solve the problem, this paper relies on boosted regression algorithm to learn variable component and influencing factors, separately. That is, multiple independent learners were combined into a strong learner to solve $G(*)$. Firstly, M^I external attributes were divided into $N-1$ groups. Let $D(h_m)_{(n)}$ be the vector of type n external attributes in period h_m. Then, training sample $\{\beta(h_m), a(h_m)\}$ satisfies:

$$\begin{aligned} a(h_m) &= WS(h_m) + TS(h_m) = G\left(\beta(h_m)\right) \\ &= \sum_{n=1}^{N-1} g^{Q}_{(n)}\left(\left(d(h_m)_{(n)}\right) + g^{P}_{(N)}\left(a\left(h_m - M^B : M^B\right)\right)\right), \\ & g_{(n)} \in G \end{aligned} \quad (6)$$

where $g(n)$, including $g^{Q}_{(n)}$ and $g^{P}_{(n)}$, is the independent learner on layer n. The independent learners were selected according to the features of historical time series and influencing factors. The historical time series $a(h - M^B : M^B)$ were treated as attributes of type N, and recorded as

$d(h_m)_{(N)}$. Then, the LS problem (2) can be optimized as:

$$\begin{aligned} G &= arg\min_{g\in G}\sum_{C}\left(a\left(h_m\right), g\left(\beta\left(h_m\right)\right)\right)^2 \\ &= arg\min_{g\in G}\sum_{C}\left(\beta\left(h_m\right), \sum_{n=1}^{N} g_{(n)}\left(d\left(h_m\right)_{(n)}\right)\right)^2 \end{aligned} \tag{7}$$

By formula (7), the LS problem was converted into the minimization of the objective of N independent learners. This paper iteratively solves each simple regression model $g_{(n.)}$ with gradient boosting algorithm. The estimation error on each layer was adjusted by a new independent learner. The synthetic model of layer $n-1$ can be expressed as:

The optimization objective on layer n can be then expressed as:

$$\begin{aligned} \min_{g\in G} G &= \min_{g\in G}\sum_{C} loss\left(a\left(h_m\right), G_{(n-1)}\left(d\left(h_m\right)_{(1)}, \right.\right. \\ & \left.\left. d\left(h_m\right)_{(2)}, ..., d\left(h_m\right)_{(n-1)} + g_{(n)}\left(d\left(h_m\right)_{(n)}\right)\right) \right. \\ &= \min_{g\in G}\sum_{C} loss\left(a\left(h_m\right), G_{(n)}\left(d\left(h_m\right)_{(1)}, \right.\right. \\ & \left.\left. d\left(h_m\right)_{(2)}, ..., d\left(h_m\right)_{(n)}\right)\right. \end{aligned} \tag{9}$$

The synthetic model G was taken as the independent variable of the loss function Loss. To further reduce the loss on layer n, the synthetic model on layer n can be updated in the functional space by gradient descent, according to the unconstrained optimization theory:

$$\begin{aligned} G_{(n)} &\approx G_{(n-1)} + \left(-\frac{\partial Loss}{\partial G}\Big| G = G_{(n-1)}\right) \\ &\Leftrightarrow \\ G_{(n)} &\approx G_{(n-1)} + g_{(n)} \end{aligned} \tag{10}$$

As shown in formula (10), the independent learner on layer n needs to fit in the negative direction of the loss function Loss on layer $n-1$ relative to G. If the loss function is MSE, the optimization objective (9) can be changed into:

$$\min_{g\in G}\sum_{C}\frac{1}{2}\left(a\left(h_m\right), G_{(n)}\left(\beta\left(h_m\right)\right)\right)^2 \tag{11}$$

To simplify the formula after derivation, the MSE was multiplied with a coefficient of 0.5, without affecting the optimization objective and model solving. The gradient of *Loss* relative to G can be expressed as:

$$\begin{aligned} &\frac{\partial Loss}{\partial G}\Big| G = G_{(n-1)} \\ &\approx \frac{\partial\left(\sum_{D}\frac{1}{2}\left(a\left(h_m\right)-G\right)^2\right)}{\partial G}\Big| G = G_{(n-1)} \\ &\approx -\sum_{C}\left(a\left(h_m\right)-G\right)\Big| G = G_{(m-1)} \end{aligned} \tag{12}$$

where $\sigma(n-1) = \Sigma_C\,(a(h_m) - G_{(n-1)})$ is the current prediction error of layer $n-1$. For the MSE loss function, the error prediction direction is the negative gradient direction of *Loss* relative to the synthetic model. The independent learner $g_{(n)} = \sigma_{(n-1)}$ on layer n and the corresponding synthetic model $G(n) = G(n-1) + g(n)$ can be obtained by fitting the prediction error $\sigma_{(n-1)}$ on the previous layer.

3. Time Series Data Fusion Algorithm for AIoT Sales

Figure 2 shows the overall architecture of time series data processing model for AIoT sales, which covers five layers, namely, hardware layer, transmission layer, data storage layer, logic layer, and application layer. To collect time series data on AP sales, the hardware layer combines data crawlers with a data mining module. The latter module is responsible for data fusion and release, configuration management, modeling and prediction, abnormality handling, and risk warning. The data crawlers are connected to the data mining module via fieldbus protocol. The data mining module aggregates the mined data, and releases them to the MQTT server, which reads the configuration. The application server then stores the collected data, predicted data, and risk warning records in the time series database. The processing of abnormal data includes padding missing data, removing noises, and removing duplicate samples. Figure 3 shows the architecture of the time series acquisition module for AP sales.

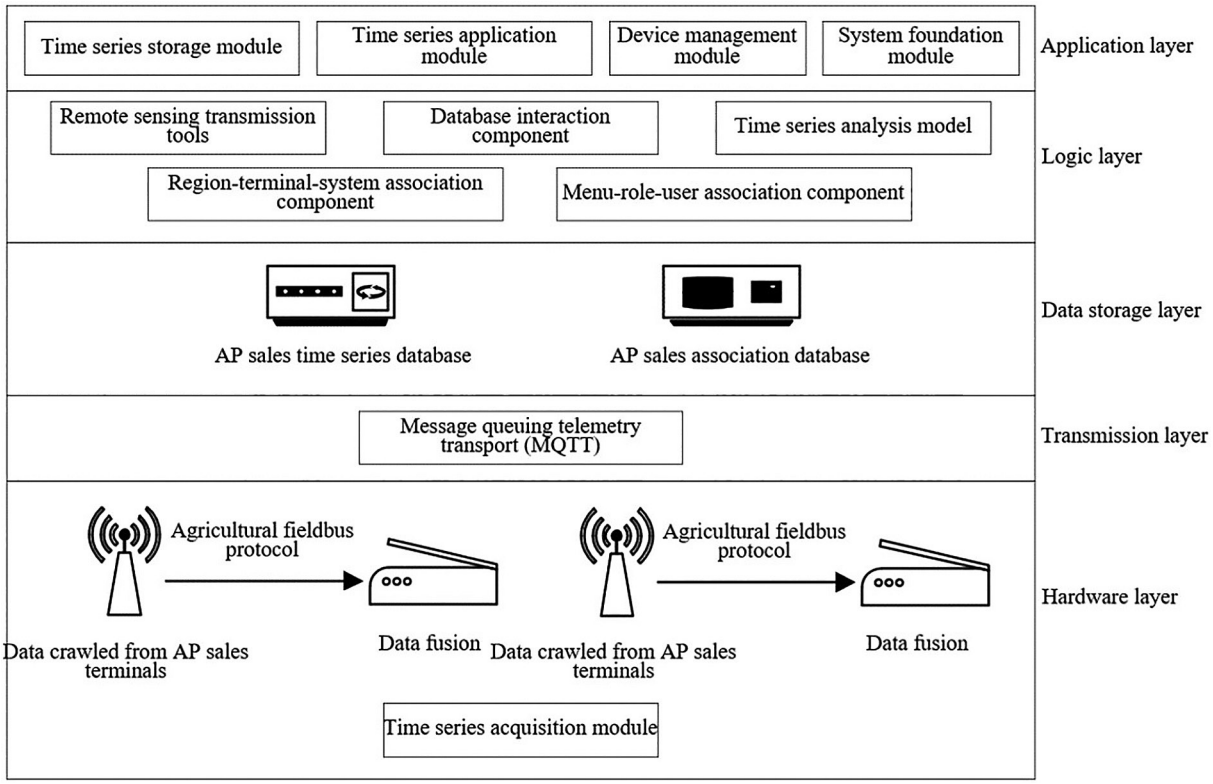


Figure 2. Overall architecture of time series data processing model for AIoT sales.

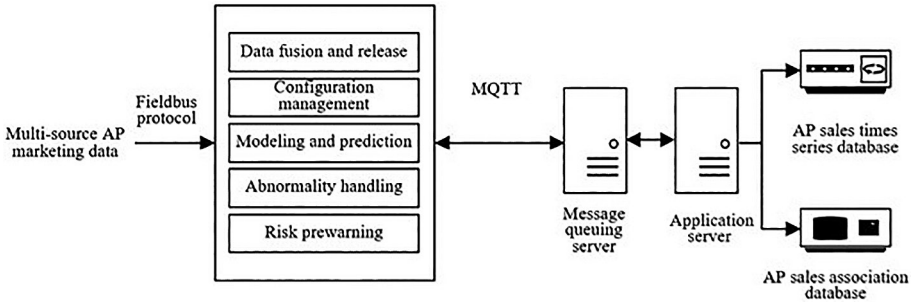


Figure 3. Architecture of time series acquisition module for AP sales.

Internet Plus brings people more channels to purchase APs. The AP marketing modes and methods are diversified by various e-commerce platforms that support online shopping of consumers, online transactions between merchants, and online electronic payment. On online transaction platforms, AP marketing data are updated and extracted in real time. However, the massive data volume makes it difficult to apply mining and analysis algorithms. This paper adopts the KF-based data fusion algorithm to process massive multi-source AP marketing data, providing samples for big-sample time series modeling of AIoT product marketing data.

In the KF algorithm, the state $A(h)$ of AP sales prediction system at time h can be given by:

$$A(h) = \lambda A(h-1) + \delta V(h) + Q(h) \quad (13)$$

The monitored AP sales $C(h)$ can be described by:

$$C(h) = FA(h-1) + U(h) \quad (14)$$

where $V(h)$ is the controlled quantity of the AP sales prediction system at time h; λ and δ are system parameters; F is the monitoring parameter. All three parameters are matrices for the multielement AP marketing data. Let $Q(h)$ and $U(h)$ be the influence of the sales process and that of monitoring, respectively; W and T be the variances of the two influences, respectively. W and T do not change with the state of the sales prediction system.

The KF is a recursive process. Based on the optimal estimate $A(h-1 \mid h-1)$ at time $h-1$ and the controlled quantity $V(h)$ of the system at time h, the estimate $A(h \mid h-1)$ at time h can be predicted by:

$$A(h \mid h-1) = \lambda A(h-1 \mid h-1) + \delta V(h) \quad (15)$$

Let λ^* be the transposed matrix of λ. Based on the covariance $cov(h-1 \mid h-1)$ of $A(h-1 \mid h-1)$, the covariance $cov(h \mid h-1)$ of $A(h \mid h-1)$ can be derived:

$$cov(h \mid h-1) = \lambda cov(h-1 \mid h-1)\lambda^* + U \quad (16)$$

The Kalman gain $ER(h)$ can be calculated based on $cov(h \mid h-1)$ and the covariance T of the monitoring influence:

$$ER(h) = \frac{cov(h \mid h-1)F'}{Fcov(h \mid h-1)F' + T} \quad (17)$$

Next, the estimate $A(h \mid h-1)$ at time h is corrected according to monitored AP sales $C(h)$, producing the optimal estimate $A(h \mid h)$ at time h:

$$\begin{aligned} A(h \mid h) &= A(h \mid h-1) + \\ ER(h)&\left[Z(h) - FA(h \mid h-1)\right] \end{aligned} \quad (18)$$

Let E be the unit matrix. Then, the covariance $cov(h \mid h)$ of $A(h \mid h)$ at time h can be calculated by:

$$cov(h \mid h) = \left[E - ER(h)F\right] cov(h \mid h-1) \quad (19)$$

Through the above prediction and correction of estimates, the optimal estimate of the current state of the AP sales prediction system can be derived from the optimal estimate of the system at the previous moment and the monitored AP sales at the current moment.

4. Big-Sample Time Series Modeling of AIoT Product Marketing Data and Long-Term Prediction

After the fusion of multi-source AP marketing data, it is the time to construct the problem of big-data time series modeling of AP marketing. Let φo be the quasi-period of an AIoT product marketing time series $\{a(h)\}^{\varphi}{}_{h=1}$; M^H and M^F be the number of inputted historical AP sales series and the number of outputted predicted future AP sales series, respectively; K^H and K^F be the input interval and output interval, respectively; R^H and R^F be the input range and the prediction range, respectively. Then, the historical monitored AP sales and predicted future AP sales can be described as $a(h-M^H: M^H: K^H)$ and $a^*(h: M^F: K^H)$, respectively. Since the long-term external factors of AP marketing are unpredictable, this paper builds up a factor-independent prediction model $s(*)$ for AP marketing time series:

$$a^*\left(h : M^F : K^F\right) = s\left(a\left(h - R^H : M^F : K^F\right)\right) \quad (20)$$

To realize long-term prediction of AP sales, K^H and K^F are set to 1. Then, $R^F = M^F$ and $R^H = M^H$. Formula (20) can be transformed into:

$$a^*\left(h : M^F\right) = s\left(a\left(h - M^H : M^H\right)\right) \quad (21)$$

where $a(h-M^H: M^H) = [a(h-M^H+1), a(h-M^H+2), ..., a(h)]$, $a(h: M^F) = [a(h+1), a(h+2), ..., a(h+M^F)]$. Next, this paper establishes the structure of a DNN for long-term prediction of big-sample time series of AP marketing data. Figure 4 shows the structure of the proposed DNN. Based on the collected big dataset of AP marketing, the proposed DNN was trained to fit the functional relationship $s(*)$ between long-term historical monitored data and predicted future time series.

The accurate long-term prediction of time series needs a long history of monitored data. To introduce long-term memory to the proposed time-varying convolutional neural network (CNN), the input historical time series should

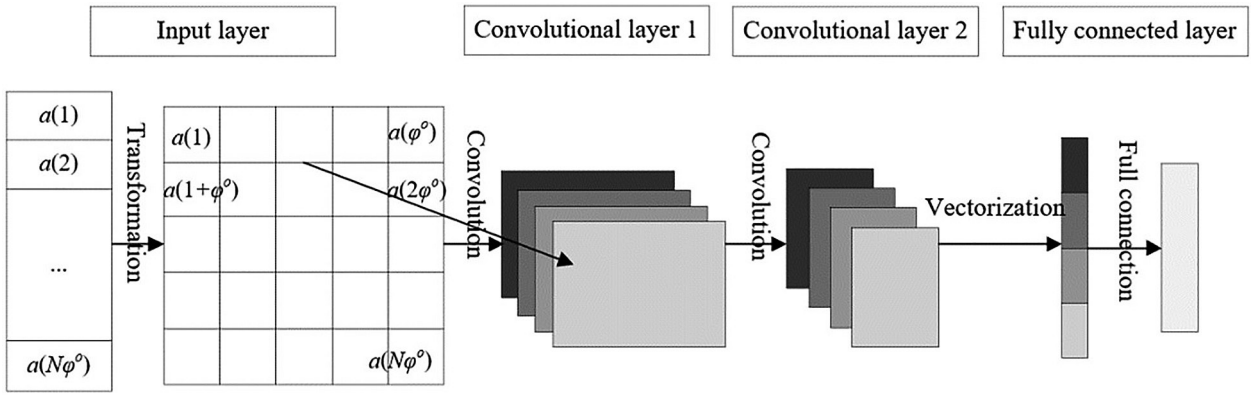


Figure 4. Structure of our DNN.

be sufficiently long. The length M^H of the input must be multiples of φ^o: $M^H = N \times \varphi^o$. Let [$a(1)$, $a(2)$, ..., $a(N\varphi^o)$] be the vector of the input historical time series at time h. By φ^o, the vector can be arranged into an $N \times \varphi^o$ matrix HA:

$$HA = \begin{pmatrix} a(1) & a(2) & \cdots & a(\phi^o) \\ a(\phi^o+1) & a(\phi^o+2) & \cdots & a(2\phi^o) \\ \vdots & \vdots & \ddots & \vdots \\ a((N-1)\phi^o+1) & a((N-1)\phi^o+2) & \cdots & a(N\phi^o) \end{pmatrix} \tag{22}$$

Suppose the k-th convolutional layer of the DNN has $M_{CK}^{J(k)}$ kernels $\omega^{J(k)}$. Let $t(k) = [t_1^{(k)}, t_2^{(k)}]$ be the vector for one translation of kernel $\omega^{J(k)}$ along the i-th dimension. Then, we have:

$$\omega^{J(k)} = \begin{pmatrix} \omega_{1,1}^{J(k)} & \cdots & \omega_{1,z_2^{J(k)}}^{J(k)} \\ \vdots & \ddots & \vdots \\ \omega_{z_1^{J(k)},1}^{J(k)} & \cdots & \omega_{z_1^{J(k)},z_2^{J(k)}}^{J(k)} \end{pmatrix} \tag{23}$$

In the DNN, the kernel size $z_1^{J(k)} \times z_2^{J(k)}$ determines the intensity of the long-term strong correlation of the time series on AP product marketing. To determine the kernel size of the first convolutional layer in our DNN, this paper calculates the autocorrelation coefficient between periods of AP marketing time series. Let LC be the linear correlation between monitoring points with an interval of i. Then, the autocorrelation coefficient ξ_i^{LC} between time series with an i-order time lag can be calculated by:

$$\xi_i^{LC} = \frac{\operatorname{cov}(a(h), a(h-i))}{\sqrt{\operatorname{cov}(a(h)), \operatorname{cov}(a(h-i))}} \tag{24}$$

Let ξ be the preset threshold and $z_1^{J(k)} \times z_2^{J(k)}$ be the kernel size. Then, the values of $z_1^{J(k)}$ and $z_2^{J(k)}$ depend on intra-period correlation and inter-period correlation, respectively:

$$z_1^{J(1)} := \left\{ e \middle| e \geq 1 : \xi_{e*\phi^O+1}^{LC} > \xi \right\} \tag{25}$$

$$z_2^{J(1)} := \left\{ e \middle| e \geq 1, 2..., \phi^O : \xi_i^{LC} > \xi \right\} \tag{26}$$

To facilitate the extraction of long-term strong correlation of time series data, a small kernel size can be designed for the kernels $\omega^{J(k)}(k \leq 1)$ in other layers. The number of kernels should double as we move to a higher level.

The output of the last convolutional layer is expanded from a two-dimensional (2D) matrix into a one-dimensional (1D) vector by a fully connected layer with a nonlinear activation function. There are M^F output nodes on the fully connected layer.

The time step has a great influence on the partial derivative of the activation function. If the time step is too long or too short, vanishing or exploding gradients will occur in the long run. In this case, the network will be unstable in training, and unable to learn early memories. This paper constructs the following periodic recursive network, such that the proposed net-

work can accurately extract the dependence of AIoT product marketing time series from the historical time series.

According to the quasi-periodicity and similar change patterns of AIoT product marketing time series, the historical monitored data were firstly divided into N subseries, each of which contains φo monitored values. Then, the N subseries were sorted by formula (22), producing an $N \times \varphi^o$-dimensional matrix.

The N subseries were linearly mapped on the fully connected layer in a unified manner. The $N \times 1$-dimensional matrix Z' outputted by the first hidden layer can be expressed as:

$$Z' = A \cdot HA^{Y(1)} = \begin{pmatrix} a(1) & a(2) & \cdots & a(\phi^o) \\ a(\phi^o+1) & a(\phi^o+2) & \cdots & a(2\phi^o) \\ \vdots & \vdots & \ddots & \vdots \\ a((N-1)\phi^o+1) & a((N-1)\phi^o+2) & \cdots & a(N\phi^o) \end{pmatrix} \begin{pmatrix} \omega_{1,1}^{Y(1)} \\ \omega_{2,1}^{Y(1)} \\ \cdots \\ \omega_{\phi^o,1}^{Y(1)} \end{pmatrix} = \begin{pmatrix} z'_{1,1} \\ z'_{2,1} \\ \vdots \\ z'_{N,1} \end{pmatrix} \quad (27)$$

Subsequently, Z' was imported to the first recursive layer. Each element in Z' corresponds to a time step: $z'(h) = z'_{h,1}{}^{(1)}$. Hence, the time step of periodic recursive network is N. Let $m^{RE(k)}$ be the number of nodes outputted on the k-th layer of the network. The output of the i-th step of the second hidden layer can be then expressed as:

$$z''(i) = AF(z''(i-1), z'(i)) \quad (28)$$

where $Z'' = [z''(1), z''(2), ..., z''(N)]$; $AF(*)$ is the activation function.

Training time and mean squared error (MSE) are selected as the metrics of model prediction effect in the present research. The M training or test sample contains M^F predicted values. Let $a(h)^{(m)}$ and $a^*(h)^{(m)}$ be the actual and predicted AP sales of the m-th sample at time h, respectively. Then, the MSE of the m-th training/test sample, and that of all M training/test samples can be respectively calculated by:

$$error^{MSE(m)} = \frac{1}{M^F} \sum_{h=1}^{M^F} \frac{\left| a(h)^{(m)} - a^*(h)^{(m)} \right|}{a(h)^{(m)}} \quad (29)$$

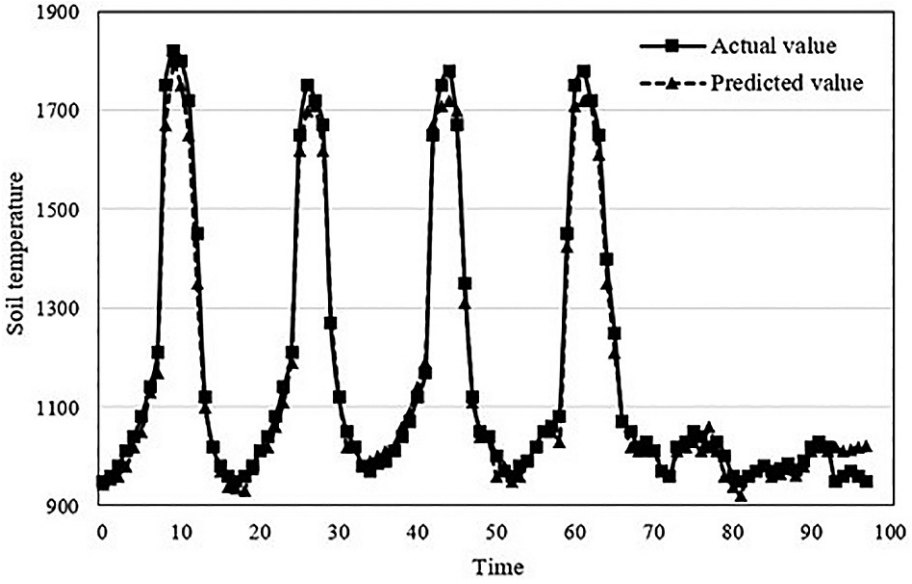
$$error^{MSE} = \frac{1}{M} \sum_{h=1}^{M} \tau^{MSE(m)} \quad (30)$$

5. Experiments and Results Analysis

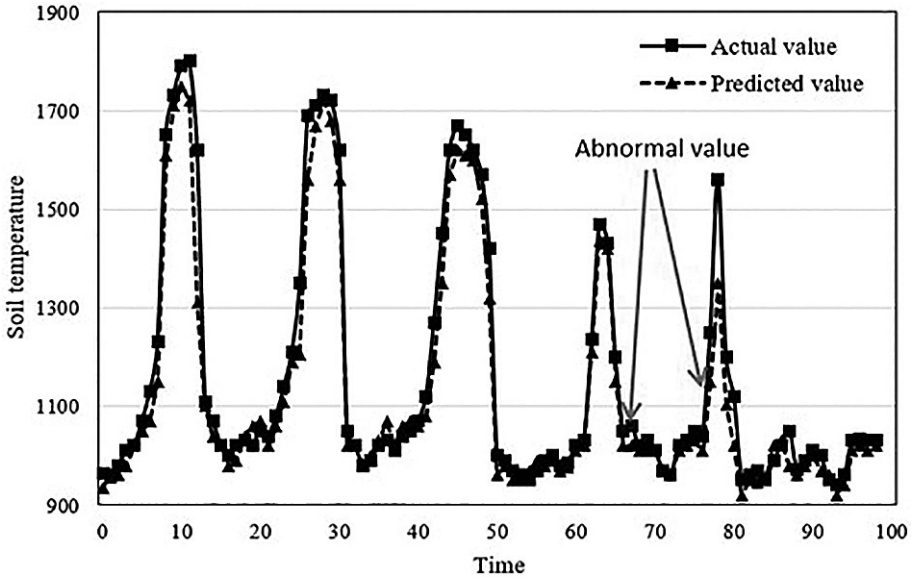
Our experiments are all based on a high-performance computer with an Intel-i7-7700K@4.2GH CPU. The experimental program was developed under Linux, specifically using the 64-bit Ubuntu 14.04, while the model was written in Python under PyCharm and Anaconda2.

The monitoring dataset on soil temperature of agricultural production depends strongly on time. The current monitoring data in the dataset are correlated with the historical monitoring data. Figure 5 compares the actual and predicted values based on monitoring data of normal and abnormal soil temperatures. It can be inferred that the prediction model, whose hidden layers contain long short-term memory (LSTM) units, did well in prediction. The monitoring dataset on soil temperature has a certain periodicity: the daytime data are generally higher than nighttime data. Note that F1= precision * recall * 2 / (precision + recall). The contrastive models include time-varying CNN, periodic recursive network, and LSTM. The results in Table 1 show that the proposed DNN performed well on time-series prediction of AP marketing in the long term, achieving a recall greater than 96%.

The monitoring dataset on CO_2 concentration of agricultural production also depends strongly on time. The features of this dataset are more unobvious than those of the monitoring dataset on soil temperature, making it more difficult to recognize the abnormal time series among this dataset. Figure 6 compares the actual and predicted values based on monitoring data of normal and abnormal CO_2 concentrations, wherefrom it can be inferred that our prediction models performed well. Table 2 lists the prediction performance indices of different models



(a) Normal monitoring data.



(b) Abnormal monitoring data.

Figure 5. Actual and predicted values based on monitoring data of normal and abnormal soil temperatures.

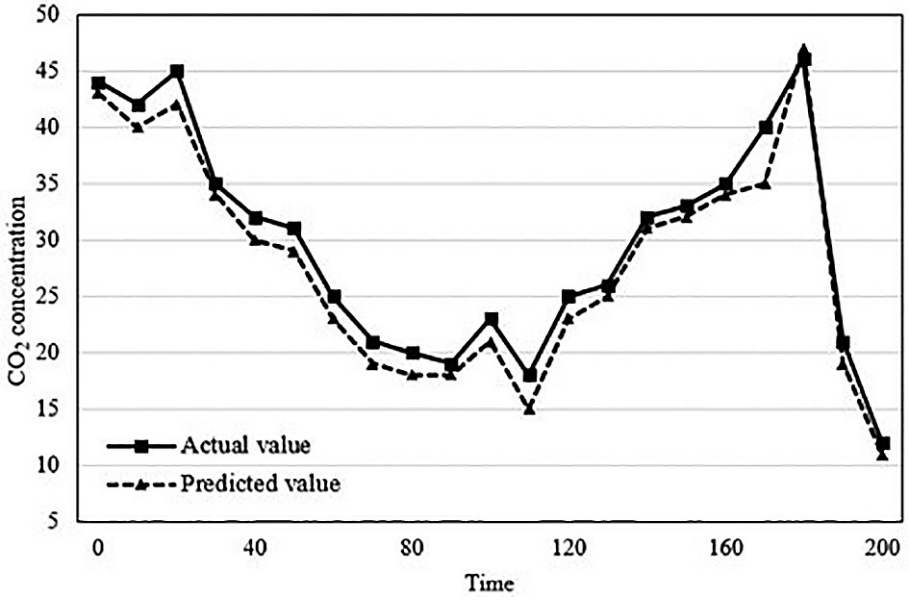
Table 1. Prediction performance indices of different models on soil temperature monitoring dataset.

Model	Accuracy	Precision	Recall	F_1
LSTM network	0.972	1	0.945	0.967
Periodic recursive network	0.907	0.853	0.936	0.889
Time-varying CNN	0.875	0.817	0.903	0.837

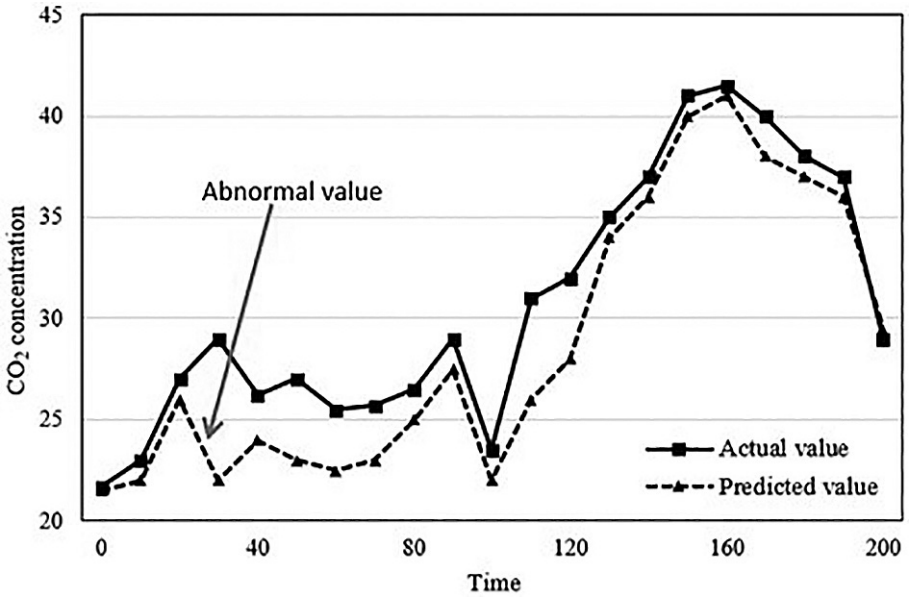
on the monitoring dataset on CO_2 concentration. The proposed models, namely time-varying CNN and periodic recursive network, performed excellently on the dataset, achieving relatively high values on all metrics.

The soil humidity monitoring dataset for agricultural production depends heavily on the time sequence. Compared with the previous two datasets, this dataset adds difficulty to feature extraction and time series prediction. Figure 7

compares the actual and predicted values based on monitoring data of normal and abnormal soil moistures, wherefrom it can be inferred that our prediction models had small prediction errors. Table 3 lists the prediction performance indices of different models on the monitoring dataset on soil moisture. The proposed models achieved ideal values on all indices, outshining the LSTM network, which cannot effectively extract the features of external factors.



(a) Normal monitoring data.

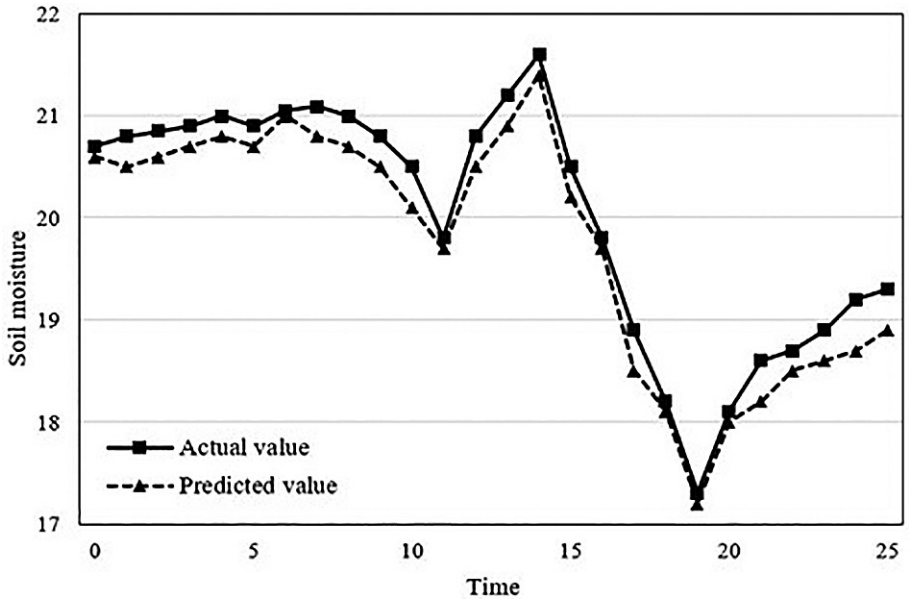


(b) Abnormal monitoring data.

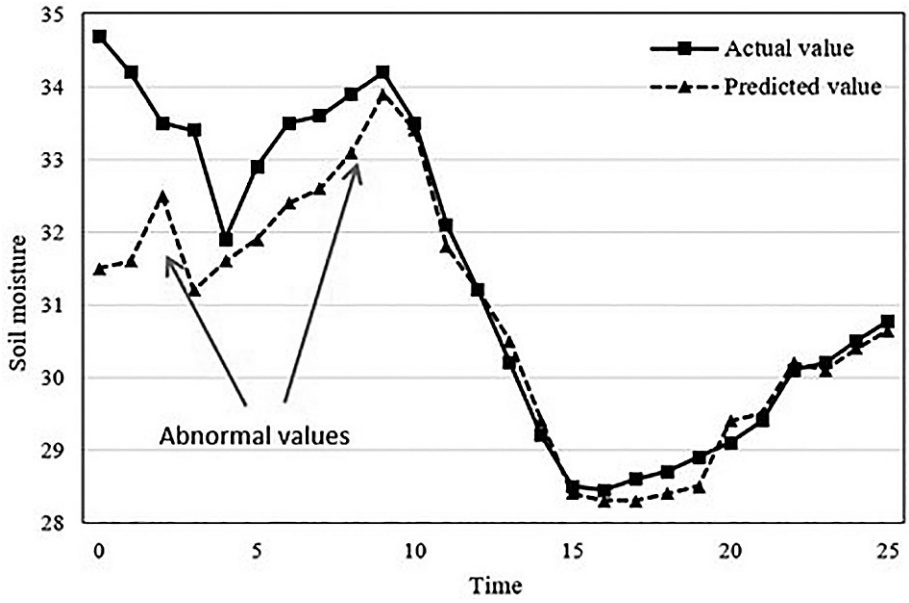
Figure 6. Actual and predicted values based on monitoring data of normal and abnormal CO_2 concentrations.

Table 2. Prediction performance indices of different models on CO_2 concentration monitoring dataset.

Model	Accuracy	Precision	Recall	F_1
Time-varying CNN	0.967	0.935	0.975	0.968
Periodic recursive network	0.872	0.863	0.872	0.894
LSTM network	0.836	0.792	0.816	0.816



(a) Normal monitoring data.



(b) Abnormal monitoring data.

Figure 7. Actual and predicted values based on monitoring data of normal and abnormal soil moistures.

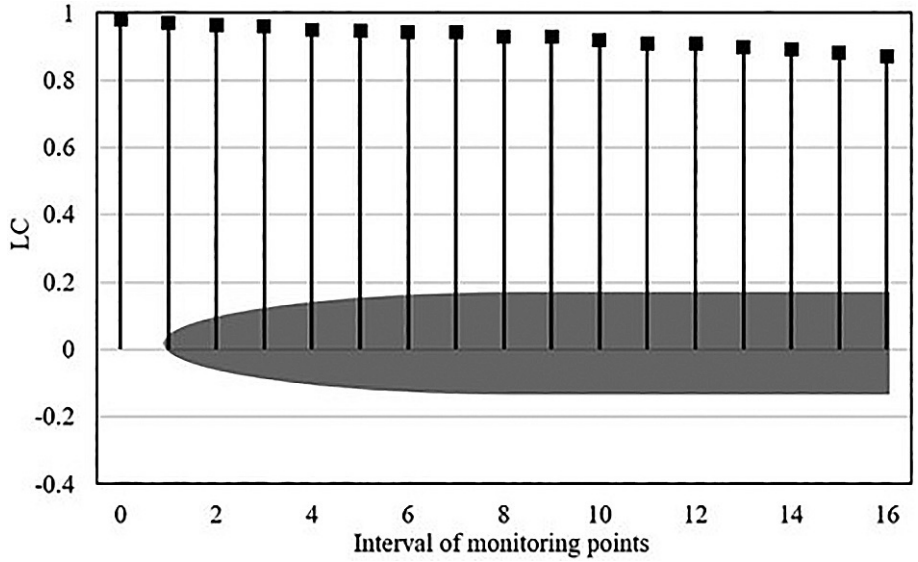
Table 3. Prediction performance indices of different models on soil moisture monitoring dataset.

Model	Accuracy	Precision	Recall	F_1
Time-varying CNN	0.975	0.931	0.968	0.846
Periodic recursive network	0.826	0.853	0.774	0.835
LSTM network	0.872	0.876	0.853	0.872

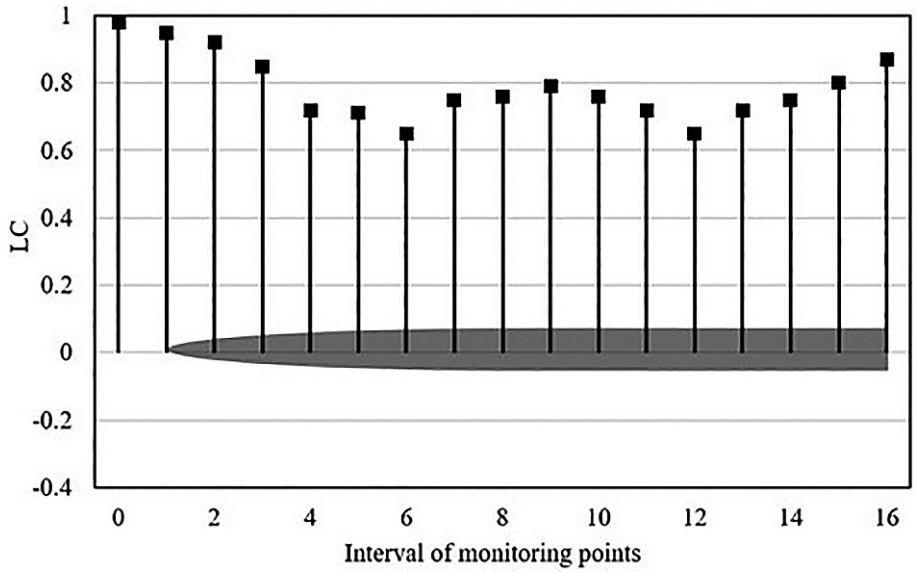
This paper analyzes the features of the AP sales datasets on vegetables and CO, respectively, and configures the proposed DNN based on the analysis results. Considering the long-term memory and quasi-periodicity of AP sales time series, the input range R^F of the sales prediction model was set to 1 year, and the quasi-period was set to 1 week, *i.e.*, $\varphi^o = 7$, and $R^F = 60 \times 7 = 420$. The historical monitored AP sales (length: 420) were sorted by the algorithm proposed in

the preceding section, and taken as the input of the prediction model. Figure 8 presents the LC values of the AP sales time series with a monitoring interval smaller than 2 weeks. Based on these LC values, the kernel size of the first convolutional layer was set to 3 × 8, and that of the other layer to 3 × 1.

To verify the performance of our prediction algorithm, the proposed time-varying CNN was employed to determine the duration of memory, and the periodic recursive network was adopted to identify the relationship between adjacent AP sales data in the same period, using one of the two activation functions: hyperbolic tangent (tanh) and linear. Finally, long-term prediction of AP sales was carried out on the AP sales datasets of vegetables and CO, respectively. This paper tests the two activation functions on the first fully connected layer of the periodic recursive network. Table 4 lists the network parameter setting and test errors. Obviously, the model prediction performance differed slightly



(a) Vegetables.



(b) Cereals and oils (CO).

Figure 8. LC values of AP sales time series with a monitoring interval smaller than 2 weeks.

under the two activation functions, suggesting the stationarity and correlation of AP sales data variation in the long term.

Table 5 compares the long-term prediction effects of our time-varying CNN, our periodic recursive network, and LSTM network. The training time of the LSTM network was 3-4 times that of the time-varying CNN, and 1.5-2.5 times that of the periodic recursive network. With the extension of prediction range, the training samples decreased gradually, the test error increased, and the training time shrank accordingly.

6. Conclusion

This paper explores the big data analysis of agricultural production, AP marketing, and influencing factors in intelligent agriculture. To realize long-, and short-term predictions, a small-sample time series model was set up for AIoT production, and a big-sample time series model was constructed for AP marketing. Besides, a KF-based data fusion algorithm was developed to fuse the massive multi-source AP marketing data. Next, experiments were carried out to compare the actual and predicted values under normal and abnormal levels of soil temperature, CO_2 concentrations, and soil moistures. The comparison shows our models achieved ideal values on all metrics, a sign of excellent prediction performance. Furthermore, the authors compared the LC values of AP sales time series with a monitoring interval smaller than 2 weeks, network parameter setting, and test errors. The results confirm the effectiveness of the proposed long-term prediction algorithm for AP sales.

Table 4. Network parameter setting and test errors.

Network configuration	Type	Training error	Test error	Training time
Input length	5 × 8	4.4	5.7	18
	30 × 8	3.6	4.1	15
	60 × 8	3.3	3.8	12
Activation function	Tanh	4.5	4.6	2
	Linear	4.2	4.5	19

Table 5. Performance of long-term prediction algorithm for AP sales.

Type	Prediction range	Algorithm	Training error	Test error	Training time
Long-term prediction	Half a year	Time-varying CNN	2.7	3.6	13
		Periodic recursive network	2.6	3.5	19
		LSTM network	2.5	2.9	55
	One year	Time-varying CNN	3.4	3.7	12
		Periodic recursive network	3.2	3.7	16
		LSTM network	4.3	4.5	45

Note: MRE is short for mean relative error.

References

[1] J. Zhang, *et al.*, "Identification of Cucumber Leaf Diseases using Deep Learning and Small Sample Size for Agricultural Internet of Things", *International Journal of Distributed Sensor Networks*, vol. 17, no. 4, 2021. https://doi.org/10.1177%2F15501477211007407

[2] A. Sánchez-Mompó *et al.*, "Internet of Things Smart Farming Architecture for Agricultural Automation", in *Proc. of the 2021 IEEE International Conference on Electro Information Technology (EIT), Mt. Pleasant, MI, USA*, 2021, pp. 159–164. https://doi.org/10.1109/EIT51626.2021.9491870

[3] W. Ren *et al.*, "A Double-blockchain Solution for Agricultural Sampled Data Security in Internet of Things network", *Future Generation Computer Systems*, vol. 117, pp. 453–461, 2021. https://doi.org/10.1016/j.future.2020.12.007

[4] T. Ojha *et al.*, "Internet of Things for Agricultural Applications: The State-of-the-art", *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 10973–10997, 2021. https://doi.org/10.1109/JIOT.2021.3051418

[5] V. K. Abrosimov *et al.*, "Agricultural Robots in the Internet of Agricultural Things", *AMA, Agricultural Mechanization in Asia, Africa and Latin America*, vol. 51, no. 3, pp. 87–92, 2020.

[6] K. Brun-Laguna *et al.*, "Using SmartMesh IP in smart agriculture and smart building applications", *Computer Communications*, vol. 121, pp. 83–90, 2018. https://doi.org/10.1016/j.comcom.2018.03.010

[7] A. Kampker *et al.*, "Industrial Smart Services: Types of Smart Service Business Models in the Digitalized Agriculture", in *Proc. of the 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Bangkok, Thailand*, 2018, pp. 1081–1085. https://doi.org/10.1109/IEEM.2018.8607270

[8] M. Ariani *et al.*, "Climate Smart Agriculture to Increase Productivity and Reduce Greenhouse Gas Emission – A Preliminary Study", *IOP Conference Series: Earth and Environmental Science*, vol. 200, no. 1, 2018. https://doi.org/10.1088/1755-1315/200/1/012024

[9] E. Effah *et al.*, "Energy-Efficient Multihop Routing Framework for Cluster-Based Agricultural Internet of Things (CA-IoT)", in *Proc. of the 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), Canada*, 2020, pp. 1–5. https://doi.org/10.1109/VTC2020-Fall49728.2020.9348608

[10] V.R. Joshi *et al.*, "Intelligent Agricultural Farming System using Internet of Things", in *Proc. of the 2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Yilan, Taiwan*, 2019, pp. 1–2. https://doi.org/10.1109/ICCE-TW46550.2019.8991914

[11] U. J. L. dos Santos *et al.*, "AgriPrediction: A Proactive Internet of Things Model to Anticipate Problems and Improve Production in Agricultural Crops", *Computers and Electronics in Agriculture*, vol. 161, pp. 202–213, 2019. https://doi.org/10.1016/j.compag.2018.10.010

[12] N. Ananthi *et al.*, "IoT Based Smart Soil Monitoring System for Agricultural Production", in *Proc. of the 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, India*, 2017, pp. 209–214. https://doi.org/10.1109/TIAR.2017.8273717

[13] Y. Matsumoto *et al.*, "Modelling and Simulation of Agricultural Production System Based on IoT Cultivated Fields Information", in *Proc. of the 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore*, 2017, pp. 354–358. https://doi.org/10.1109/IEEM.2017.8289911

[14] K. Wongpatikaseree *et al.*, "Developing Smart Farm and Traceability System for Agricultural Products Using IoT Technology", in *Proc. of the 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore*, 2018, pp. 180–184. https://doi.org/10.1109/ICIS.2018.8466479

[15] M. Lee *et al.*, "Agricultural Production System Based on IoT", in *Proc. of the 2013 IEEE 16TH International Conference on Computational Science and Engineering, Sydney, Australia*, 2013 pp. 833–837. https://doi.org/10.1109/CSE.2013.126

[16] Q. Wang and X. Yang, "Research on IOT based Special Supply Mode of Agricultural Products", in *Proc. of the 2014 International Conference on Mechatronics, Electronic, Industrial and Control Engineering*, 2014, pp. 1715–1718. https://dx.doi.org/10.2991/meic-14.2014.392

[17] L. Mo, "Study on Supply-Chain of Agricultural Products Based on IOT", in *Proc. of the 6th International Conference on Measuring Technology and Mechatronics Automation, Zhangjiajie, China*, 2014, pp. 627–631. https://doi.org/10.1109/ICMTMA.2014.153

[18] P. R. Harshani *et al.*, "Effective Crop Productivity and Nutrient Level Monitoring in Agriculture Soil Using IoT", in *Proc. of the 2018 International Conference on Soft-computing and Network Security (ICSNS), Coimbatore, India*, 2018, pp. 1–10. https://doi.org/10.1109/ICSNS.2018.8573674

Received: May 2022
Revised: July 2022
Accepted: July 2022

Contact address:
Jianfeng Cheng
Economics and Management School
Wuhan University
Wuhan
China
e-mail: 2014101050132@whu.edu.cn

Jianfeng Cheng is currently pursuing a Ph.D. degree at Wuhan University. He received his master's degree from Sun Yat-sen University, in 2009. He is currently a specialist in his industry and has excellent working experience in practical operation management for over decades. His research interests include marketing and data mining, and he keeps tracking the trends in applied computing.