SAMPLING METHODS USING REMOTE SENSING AND GLOBAL POSITIONING SYSTEM FOR CROP ACREAGE ESTIMATION AT - NATIONAL SCALE IN CHINA

Quan Wu [a], *, Li Sun [a]

a CAAERP, Chinese Academy of Agricultural Engineering Research and Planning, 100125, No. 41 Maizidian Street Chaoyang District, Beijing
– wuquan95@tom.com,- sunli0618@163.com

Commission VII, WG VII/7

KEY WORDS: Sampling, Sampling unit, Stratified sampling, RS, GPS, Agricultural condition, Acreage

ABSTRACT:

Sampling methods have to be applied in agricultural condition monitoring at national scale. When selecting sampling methods, more attention needs to be paid not only to objects monitored, but also to conditions monitoring such as money and time, etc. Stratified sampling method together with remote sensing has been successfully applied to main crops acreage monitoring for many years in China. Acting as the important complement for RS, ground sampling method system has been established in estimating main crop acreage. At a large enough scale, high efficiency and operability are key points in agricultural condition monitoring. Obviously, proper sampling methods have to be employed in agricultural condition monitoring using remote sensing at national scale because the total overcast investigation is not only unnecessary but also impossible. Thereby, it is very important to choose appropriate sampling methods and sampling methods need to be adapted to different tasks and crops. Crop acreage is very important information in agricultural condition monitoring using remote sensing or ground sampling methods. Stratified sampling using RS and ground random sampling methods are adopted by the Chinese RS Application Centre of Ministry of Agriculture in getting the acreage of winter wheat, corn, soybean, rice and cotton.

1. THE BACKGROUND

1.1 Agricultural Condition

Agricultural production is significant to any country, especially to China because of the great number of rural population. In 2006, there were more than 470 million people among 900 million country people working on farmland and the gross value of agricultural production was more than RMB 2400 billion. Therefore, it is of great importance for local and central governments to get accurate agricultural production information in time. Agricultural production information, which is called agricultural condition in China, includes some items as follows:

1. Crop Acreage
2. Crop Yield
3. Natural Disaster (Flood, Drought, Insects, etc.)
4. Growth condition

Subject to the Ministry of Agricultural of China, Remote Sensing Applications Centre (RSAC) has been working on the above things as operating tasks and research projects for several years. In this article, methods for crop acreage estimation will be discussed.

1.2 Crop Acreage

Crop acreage, which means planting area of crop, is one of the most important agricultural condition information. Crop acreage is directly and closely related to crop's total yield. Because crop yield per unit usually varies in the range estimated in normal year, crop acreage varies comparatively greater than crop yield per unit does, so is total yield, hence crop acreage information is even more important than crop yield per unit. There are many kinds of crops planted in China, but RSAC pays more attention to the main crops such as wheat, corn, cotton, soybean and rice, etc. Due to the natural condition and planting habits, these crops are mainly centralized in about 15 provinces, but the planting time varies from province to province. Table 1 shows the planting regions and time for these main crops. For the purpose of getting these crops acreage data every year, RSAC has to select appropriate method to estimate.

Crops	Distributed area	Planting time
Spring wheat	Northwest and Northeast China	Mar.-Apr.
Winter wheat	Middle and East China	Sep.- Oct.
Corn	Northeast, Middle and East China	Apr.- Jun.
Cotton	Northwest, Middle and East China	Apr.
Soybean	Northeast ,Middle, East China	Apr. - May
Rice	Northeast, Southeast and South China	Jan.-Mar., Jun.-Aug.

Table 1. Crop planting area and time in China

1.3 Methods of getting crops acreage data

Generally, there are two ways of getting crops acreage data: one is from government statistics and the other is general survey

* Corresponding author.

data provided by the operating department subordinated to the government. In order to obtain the data of crops acreage every year in good time, RSAC has to consider two factors as the operating department, which are time and money. Since the spatial range monitored is very large and the kinds of crops are so many, it is impossible to investigate the overall fields either by RS or by ground survey using GPS. Therefore, RSAC chooses sampling method using RS and GPS.

2. OTHER WAYS OF GETTING CROP AREA DATA

2.1 In European Union (EU)

Monitoring Agriculture with Remote Sensing (MARS) is a project facing European in order to obtain crop yield information constituted by European Union Committee (Liu, 1999). It is a kind of three-stage sampling based on unsupervised classification (Duda, 2002) using multitemporal RS data (Panigrahy, 1997). The first-stage sampling unit composed of 60 sites is designed square with the side length of 40 kilometres. There are 16 component parts in each site and each part has 40 sampling points. Because unsupervised classification method is used with RS data to cover the third-stage sampling points 5-6 times every year, all crops in sampling units are recognized and then the acreage and yield will be worked out after statistical calculation from the third-stage sampling points to the first-stage sampling units.

2.2 In America

In America, the prediction of total crop yield is acquired from crop acreage and crop yield per unit. The crop acreage data had been gotten by June Agricultural Survey (JAS) (Hu, 2002). Two different sampling units used by JAS were area frame covering America and name list frame. The name list was composed of registered farmers. Every year about 2400 investigators contacted more than 120 thousand farmers in the first two weeks of June in order to get crops acreage data.

2.3 In China

The operating prediction of crops acreage is mainly provided by RSAC. RSAC adopts two methods to obtain the acreage of main crops such as wheat, corn, cotton, soybean, rice, etc. One method is stratified sampling using RS and the other is ground random sampling using GPS. Stratified Sampling Method (SSM) using RS is the major one that works as illustrated by the flow chart below (Chen, etc., 2000).

To use SSM, the first step is to select the sampling units according to the surveyed crop and to order appropriate RS images covering all sampling units in the region. The second step is to discriminate the crop with RS data and get the area data of the crop distributed in the sampling units. The last step is calculating the result with SSM.

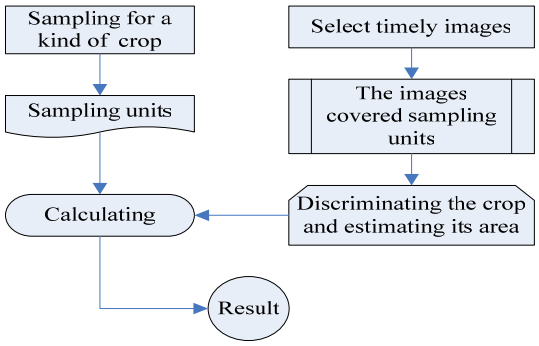

Figure 1. Flow chart of getting crop area by SSM

3. SAMPLING METHODS

RSAC selects the sampling methods that include stratified sampling in spatial regions with RS and Ground Random Sampling (GRS) with GPS.

3.1 Stratified sampling using RS and GIS

The aim of sampling is to estimate the total quantity of the object such as crop area since the total quantity is too large to count and it is of no necessity to survey entirely. Stratified sampling is a kind of methods like random sampling, multi-stage sampling, etc. It is mainly applied to such case as that the target individuals are very different from one another but have quantitative attributes

Selecting sampling unit: Sampling unit has two key points: quantity and accuracy. Meanwhile, it should be convenient for operating and calculating reasons. RSAC selects two kinds of sampling units: one is the polygon of administrative counties, and the other is the quadrangle frame of relief map on which the scale used is 1:50,000 or 1:25,000, and longitude difference one quarter degree and latitude difference one sixth degree (Jiao, 2002). The difference between the two sampling units is that the former is easier for statistics analysis but the quantity is too small to satisfy the demand of sampling some crops because the areas of sampling units are apparently different among the polygons of administrative counties, the latter is just opposite.

3.1.1 Layers and layer amount: In stratified sampling survey, layer amount has influence on the effect of sampling. The proper amount of layers is related with the characters of number of sampling units. The appropriate amount of layers will minimize the population variance of total layers and the sampling cost. During a certain threshold interval, adding layer number could commonly lower population variance, but increase the workload. In an attempt to balance the effect of sampling and the expense of survey, the number of six is confirmed the maximum layer amount in sampling survey of crops area according to many tests done by RSAC.

The so-called layer is a kind of data set based on sampling units. There are obvious differences in sampling size among the layers. In survey of crop samplings, the sampling units that belong to a certain layer are generally distributed relatively centralized. The map below shows the distribution of six layers of Chinese late rice of 2007, on which sampling units is the quadrangle frame of relief map with the scale 1：50000. There are totally 5340 sampling units covering 15 provinces of China in this map. It is

obvious from the map that the sampling units that belong to the same layer are relatively centralized in the area.

3.1.2 Selecting background data for stratified sampling: As soon as the background data for stratified sampling is selected, sampling units are also determined. RSAC selects two kinds of background data to do stratified sampling. One is multiyear statistic data collected by local governments and the other is the latest land use data in vector format.

Processing statistic data: Because statistic data mainly comes from local governments, borderlines of administrative units such as counties generally have to be selected to act as sampling units when the statistic data is used to stratify sampling. The main step is to average the multiyear data of each county, whether making ascending or descending data array totally based on the method of stratified sampling. Stratified sampling steps should correspond to the following presentation. Each county should be marked with layer sign.

Processing vector data: When the land use data is selected as background data to stratify sampling units, there are some things to consider. Firstly, it is necessary to select sampling unit such as the frame of relief map used by RSAC. Secondly, it is a key step to combine the vector data using land use dada with frame data of relief map in GIS, shown below as an example. Thirdly, it is to calculate surveyed crop area such as rice, soybean, corn, cotton, or wheat, etc, that distributes in every frame of relief map. The final step is to sort in ascending or descending order according to the crop area of every unit if necessary.

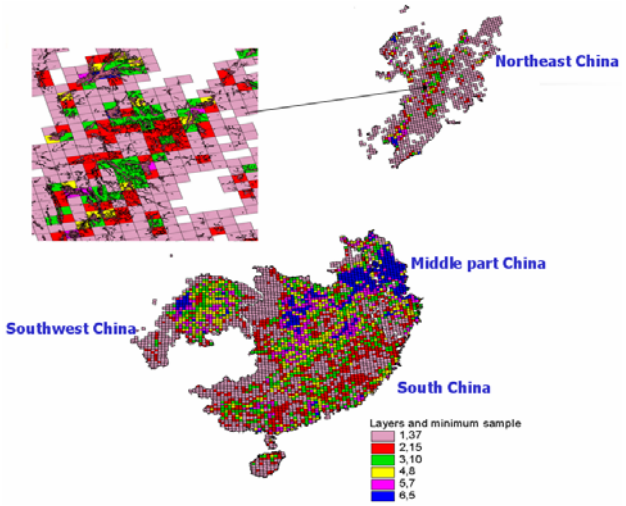

Figure 2. Layers distribution of late rice of 2007 in China.

3.1.3 Methods of stratified sampling: After the layer amount is determined, the next step is to stratify. Two stratify ways used by RSAC are called Frequency Accumulation Means (FAM) and Systematic Clustering Means (SCM).

FREQUENCY ACCUMULATION MEANS (FAM)

RSAC mainly uses this method to stratify sampling units in area survey of cotton, rice, soybean, etc. The essence of the method includes dividing data into groups according to a certain step length between two groups in ascending or descending data array, calculating the amount of sampling units of each group, which is called frequency, accumulating frequency and square root of frequency of each group, getting the total accumulating value of square root by adding up square roots of frequency of total groups, obtaining the step length between two layers through dividing the total accumulating value of square root by layer amount such as six, which is an equal step length method. The thresholds of layers have been shown in table 2 and they will be used for segment point marking each layer.

When the background data is processed to be either in ascending or descending data array, the following steps stay the same whether it is statistic data or land use data, as shown below by the example of the process of stratified sampling of early rice of 2007 in China. Table 2 shows the key points: the first column on the left side is codes list of frame of relief map, the second is the crop area distributed in frames of relief map.

codes of relief map	Area of rice(ha.)	Frequency $f(y)$	Accumulating $\sqrt{\sum f(y)}$	Thresholds of layers	Layers
7500123	0.04	0	0.00		1
……	……	……	……		
6491261	2.3	7	2.65		1
……	……	……	……		
6490013	4058.72				1
7500252	4060.22	3	9286.06	9288.83	1
7500314	4063.05	1	9316.52		2
……	……	……	……		
7490053	6547.84				2
7491214	6550.25	2	18570.81	18577.65	2
8501173	6554.17				3
……	……	……	……		
6500141	9534.13				3
7491283	9537.54	2	27842.35	27866.48	3
7500861	9561.18	1	27884.69		4
……	……	……	……		
7490562	13310.9	1	37078.12		4
9501293	13322.8	1	37123.62	37155.30	4
7500283	13328.1	1	37169.14		5
……	……	……	……		
8491412	19009.5	1	46379.36		5
8501382	19027.2	1	46427.30	46444.13	5
8491174	19033.9	1	46475.24		6
……	……	……	……		
8500182	40091.7	1	55732.95	55732.95	6

Table 2. The process of stratified sampling

Grouping: Dividing data into groups according to a certain step length between two groups. RSAC has selected 4.5 hectare as the step length between two groups, which is about one ten-thousandth of a frame area, produced 1550 groups from the sampling units of 2497, marking at the position of each group before the threshold of next group in the data array using code, which position can be called Position of Group Threshold(PGT), correspondingly producing a new column which is omitted in table 2.

Counting the frequency: Counting and marking the number of sampling units of each group at PGT, which is called frequency listed in the third column on the left side of table 2, signed with $f(y)$.

Accumulating the frequency: A new column omitted in table 2 was correspondingly produced, which was used to store the data of accumulating frequency when the frequency of each group was accumulated and then marked at each PGT from top to bottom of this data array. the data of accumulating frequency was signed with $\sum f(y)$.

Accumulating $\sqrt{\sum f(y)}$: A new data array was calculated and marked by accumulating the square root of the data of accumulating frequency at each PGT, which is listed in the fourth column on the left side of table 2.

Thresholds of layers: The distance datum between two layers can be calculated using the last accumulating data of the square root of the data of accumulating frequency which was shown at the bottom of the third column on the right side of table 2 divided by layer amount such as six. This distance datum which is 9288.83 in table 2 was used to partition the data array of the third column on the right side of table 2 and mark at six PGT listed in the second column on the right side of table 2.

Marking layers: Every sampling unit can be marked with the sign of layers when the thresholds of layers are confirmed. In the above table, total sampling units were partitioned to six regions according to the thresholds of layers, and every region was marked with numerals as codes of layers. These signs of layers were listed in the first column on the right side of table 2.

SYSTEMATIC CLUSTERING MEANS (SCM)

It is reasonable to stratify sampling units using systematic clustering for reference. The result of clustering is similar to that processed by FAM although there is only one parameter used to cluster, that is crop area such as paddy rice of sampling units. The main steps will be introduced corresponding to the process of stratified sampling of early rice of 2007 in China.

Preparing data: The data used to cluster need not to be sorted ascending or descending. It only needs to be transformed into appropriate format for statistical software such as SPSS.

Clustering: It is easy to cluster when the background data is imported using statistical software. RSAC also selects the number of six as the number of clustering in stratified sampling of early rice of 2007 in China. The clusters of six were used as layers of six.

3.1.4 Calculating the least samples: The following equations should be used to calculate the least samples when the job of stratified sampling has been done (Chen, etc., 2000).

$$\overline{Y_h} = \frac{1}{N}\sum_{i=1}^{N_h} Y_{hi} \tag{1}$$

$$S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h} (Y_{hi} - \overline{Y_h})^2 \tag{2}$$

where $\overline{Y_h}$ = the average value of the collectivity of the h layer

N = the amount of the collectivity

N_h = the amount of the collectivity of the h layer

Y_{hi} = the total crop area of the i unit of the h layer

S_h^2 = the summation of variable error of the h layer

$$\overline{y_h} = \frac{1}{n}\sum_{i=1}^{n_h} y_{hi} \tag{3}$$

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h} (y_{hi} - \overline{y_h})^2 \tag{4}$$

where $\overline{y_h}$ = the average value of the samples of the h layer

n = the total amount of the samples

n_h = the amount of samples of the h layer

y_{hi} = the crop area of the i unit of the h layer

s_h^2 = the variable error of the samples of the h layer

So, the least samples amount of stratified sampling is determined by the follow formula:

$$n = \frac{\sum_{h=1}^{L} \frac{N_h s_h^2}{W_h}}{V + \sum_{h=1}^{L} N_h s_h^2} \tag{5}$$

Where n = the total amount of samples

$W_h = N_h / N$

L = the amount of layers

V = the variable error of estimating value

The others are as same as the above

Distribution according to proportion:

$$n = \frac{n_0}{1 + n_0 N} \tag{6}$$

Where $n_0 = \frac{N\sum_{h=1}^{L} N_h s_h^2}{V}$

The others are as same as the above

The surveying accuracy signed δ is set to a certain value such as 0.95, and the reliability signed $(1-\alpha)$ is also set to 0.95 or other value, the value of V should be calculated according to the follow formula:

$$V = (d/t)^2 \tag{7}$$

where $d = (1-\delta)Y$, Y = the total value of the collectivity

$t = \mu_{\alpha/2}$ when the number of samples is bigger than 45

RSAC has used the above formulas to calculate the least samples in stratified sampling of early rice of 2007 in China using FAM and SCM, and the result is listed in table 3. Based on the methods of stratified sampling, the sampling proportion is different between the FAM and the SCM. The former is 0.0377, and the latter is 0.0231. Sampling with FAM needs the least samples of 94, and with SCM only 58. Sampling with SCM seems better than with FAM at the point of sampling proportion. However, the numbers of sampling units distributed to six layers change more violently with SCM than FAM, which leads the least samples of six layers to the same situation.

layer	total units		the least samples	
	by FAM	by SCM	by FAM	by SCM
1	927	1014	35	24
2	506	755	19	18
3	358	60	14	1
4	280	394	11	9
5	227	201	9	5
6	199	73	8	2
total	2497	2497	94	58

Table 3. The result of calculating the least samples

3.1.6 Sampling using RS and GIS: When the job of stratified sampling has been done, the next step is to show the layers in GIS as figure 2. Then, some RS images covering the sampling units should be ordered according to the least samples of every layer, and imaging date should be in early planting days of the surveyed crop. The area covered by images ordered should not be less than the area of the least samples of every layer.

Because of the clouds covered in the paddy region of three seventh during the period of surveying early rice of 2007 in China, RSAC only ordered 36 SPOT images in stratified sampling. On the basis of this status, the surveying collectivity was adjusted and sampling units was stratified once again. The new number of the collectivity is 1699, and the summation of least samples of six layers is 74 using FAM, it is 56 using SCM. However, the total number of sampling units is 201 in the actual application.

3.1.7 Estimating the collectivity: Interpreting RS images covering the sampling units is an important step before estimating the surveying collectivity (Thomas, 2002). The quality of interpretation is closely related to the result of surveying (Yang, 2002). The key point of interpretation is to discriminate the crop and get the crop area of each sampling unit covered by RS images. After the total sampling units covered by RS images have been interpreted, the total area of surveying crop is to be estimated using the following formula:

$$\hat{Y} = \sum_{h=1}^{L}\sum_{j=1}^{N_h}\left(\frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}\right) \tag{8}$$

Where y_{hi} = the crop acreage of unit i of the h layer

N_h = total number of sampling units of the h layer

$\hat{Y}$ = estimate value of total area of the collectivity

L = total number of the layers,

$h = 1, 2, \ldots, L$

n_h = the amount of sampls of the h layer

RSAC used two continuous years' RS data to calculate the crop area variation rate based on SSM. Using FAM, the variation rate of area of early-rice covered the four-sevenths paddy fields of China is –2.48% from 2006 to 2007 and the confidence interval is from –8.91% to 4.41% while the confidence is 95%. Using SCM, the variation rate is -2.30% and the confidence interval is from –9.51% to 5.49% while the confidence is 95%.

3.2 Ground random sampling using GPS

Ground Random Sampling (GRS) using GPS is an independent method adopted by RSAC. On one hand, using GRS can make up for the lack of RS such as images covered by clouds. On the other hand, ground sampling can provide independent information of agricultural condition such as crops area, and crops geographical position information, which provides reference to the interpretation of RS images.

3.2.1 The shape and size of sampling unit of GRS: The sampling units of GRS can be called sampling frames. The sampling frames are designed as polygons that are located on farmland by RSAC (Chen, 1990). The polygons are mainly made up of natural borderlines coming from land cover such as road, dyke, ribbing, etc. Each polygon area is about 25 hectare. The structure is shown below. The two sampling frames used by RSAC to survey the area of rice are distributed in Guangdong province of China. In the map the codes indicate different land cover. The code of 1100 indicates paddy field, the code of 2000 indicates fallow, the code of 8001 indicates dykes, the code of 7000 indicates roads, the code of 1800 indicates vegetables, the code of 3000 indicates garden plots, and the code of 1901 indicates other plots of crops.

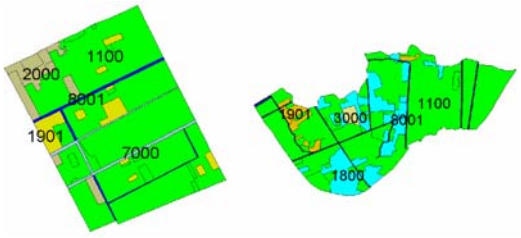

Figure 2. The structure of sampling frames

3.2.2 Getting the data of crops area: Using GPS to get spatial information attribute of polygons. The proportion of surveying crop such as rice based on sampling frames is easily calculated in GIS, which is used as sampling unit data. RSAC used two continuous years' data of GRS to calculate the variation rate of crop area based on statistical rules at a certain surveying region such as administrative district.

3.2.3 Sampling accuracy and samples amount: It is obvious that increasing samples amount can improve the sampling accuracy. However, increasing samples amount will definitely increase the cost of surveying. Thus it is necessary to balance the cost and the accuracy of surveying. The formula used to express the relation between samples amount and sampling accuracy is shown as below (Wu, 2004):

$$\Delta \geq \frac{2\sigma}{\sqrt{n}} Z_{\alpha/2} \tag{9}$$

Where σ = standard deviation
n = samples amount
Δ = sampling accuracy
Z = the parameter of standard normal distribution
$1-\alpha$ = confidence level

In this formula, the standard deviation can be replaced with the variable error of samples via testing in advance. When the sampling accuracy is fixed on a certain value such as 0.95, the least samples amount can be calculated.

4. CONCLUSION

Using RS and GIS, stratified sampling method has been successfully applied to main crop area monitoring for many years at national scale in China. RSAC uses two kinds of methods to stratify: one is called Frequency Accumulation Means (FAM) and the other is System Cluster Means (SCM). The two methods have different characters. With GPS, Ground Random Sampling (GRS) has also been adopted as complement to RS when estimating the variation rate of main crop' area in China.

ACKNOWLEDGEMENTS

The data used for stratified sampling and calculating crops area is produced by RSAC. With deep gratitude we want to give our thanks to all the members of RSAC.

REFERENCES AND/OR SELECTED BIBLIOGRAPHY

Chen, Z., etc., J.,2000. Sampling and scaling scheme for monitoring the change of winter wheat acreage in China. *Transactions of the Chinese Society of Agricultural Enginerring*, 16(5), pp. 126-129.(in China).

Duda T., etc., J.,2002. Unsupervised classification of satellite imagery: choosing a good algorithm. *Remote Sensing*, 23(11), pp. 2193-2212.

Jiao, X., etc., J.,2002. Design of sampling method for cotton field area estimation using remote sensing at a national level. *Transactions of the Chinese Society of Agricultural Enginerring*, 18(4), pp. 159-162.(in China).

Liu, H.,J.,1999. MARS project of European Union and plan of China using RS. *Chinese Journal of Agricultural Resources and Regional Planning*, 20(3), pp. 55-57.(in China)

Panigrahy S., etc., J.,1997. Mapping of crop rotation using multidate Indian remote sensing satellite digital data. *Photogrammetry & Remote Sensing*, 52, pp. 85-91.

Wu, Q., J.,2004. Influence of small features on crop area estimation at a nation scale using remote sensing and a double sampling method. *Chinese Journal of Agricultural Resources and Regional Planning*, 20(3), pp. 130-133.(in China)

Yang X., etc., J.,2002. Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area. *Remote Sensing*, 23(9), pp. 1175-1798.

Chen, S.,etc, M.,1990. *Geography Analysis of Remote Sensing.* Beijing, Mapping publishing, pp. 147-167. (in China)

Hu, R., etc, M.,2002. *Application of RS and yield estimation of Agricultural crops in European Union*. Beijing, Weather publishing, pp. 34-59. (in China)

Thomas M. Lillesand, etc.,2002. Remote sensing and imagery interpretation. New York, John Wiley & Sons, Inc, pp. 576-586.